Meta-Coherence: A Cross-Domain Empirical Specification for Validating Induced Coherence Weights Without Correspondence Truth

Benjamin James

May 1, 2026

Abstract

Coherence-Weighted Information Theory and the induced-weights program that grounds it require empirical validation across the cross-domain regime where their universality claims live. Adaptive Realism and Synthetic Epistemology forbid correspondence-truth validation at the foundational level. The real is what persists across recursive coherence arbitration, and external ground-truth oracles violate the framework's ontology before any measurement is taken. I specify a meta-coherence validation architecture that operates inside the framework's commitments. Three within-domain signatures of self-consistent measurement are required jointly: cross-philosophy convergence (R2) across estimator classes with genuinely different inductive biases, persistence prediction (R1) grounding induced weights in Adaptive Realism's persistence criterion, and intervention asymmetry (R3) operationalizing the substrate-symbol distinction. A fourth condition, M5, carries the domain-translation invariance commitment as a structurally distinct admissibility gate sequenced before the within-domain matrix. The program tests these conditions across three domains spanning a substrate-stream transparency gradient, including a hidden semi-Markov source with structured emissions, Pfam protein families, and FOMC statements. An eight-cell outcome interpretation matrix binds in advance what each result configuration means for specific framework claims. The paper specifies the program; running it is the next layer.

Keywords: coherence information theory, induced coherence weights, meta-coherence validation, adaptive realism, synthetic epistemology, symbolic lock-in, cross-domain admissibility, pre-registration, intervention asymmetry, recursive coherence

The Validation Problem That Cannot Be Solved by Validation

After *Formal Foundation of Induced Coherence Weights*, the weight function *w* is no longer arbitrary. The derivation pipeline from observed state streams to coherence proxies, from coherence proxies to symbol-level relevance, and from relevance to bounded weights is specified end to end. The admissibility commitments of boundedness, monotonicity, coarse-graining consistency, recursive stability, empirical derivability, and domain-translation invariance bind the weight functional as constraints it must satisfy or fail. Predictions follow with explicit structure; coherence-weighted compression preserves coherence-bearing structure under distribution shift, weight rankings remain stable under benign coarse-graining, declines in the coherence proxy precede regime transitions, and constraint-relief interventions measurably outperform purely interpretive ones. The program is, in principle, falsifiable. What it is not yet, is tested.

One residual seam was named at the close of that prior work and left open. The induced weights inherit the inductive biases of whatever compression estimator K computes the proxy from which they are derived. If coherence weighting depends on the idiosyncrasies of K, the universality claim collapses into estimator artifact, and the formal results, coherence-weighted capacity and selective compression bounds, become statements about the chosen compressor rather than about the system being compressed. The response previously proposed was structurally correct; reframe invariance as robustness, require convergence across reasonable estimator classes up to monotone transformation, and treat the seam as a research program rather than an evasion. That response is necessary. It is also insufficient on its own. It tells us what convergence would look like if it held; it does not specify the estimator class across which convergence would actually carry weight as evidence rather than as artifact of philosophical agreement among the chosen compressors.

The natural next move, in any other empirical context, is to validate against ground truth. Pick a domain where structural features are known, such as catalytic residues annotated by enzymologists, regime breakpoints documented by economic historians, or narrative arcs tagged by literary critics, and check whether high-*w* features track those features. If they do, *w* has been validated. If they do not, *w* has been refuted. This is how empirical claims about latent structure are usually discharged; an authoritative external label adjudicates, the framework either matches the label or fails to, and the verdict is registered.

Adaptive Realism forecloses this move at the level of what ground truth can mean. The real, in this framework, is not an external bedrock against which symbols are compared. It is what persists across recursive coherence arbitration, that which survives, under perturbation, in the substrate's

own ongoing negotiation with entropy. There is no oracle layer beneath the symbol stream from which a verdict can be drawn, because the framework denies that such a layer exists as a static absolute. Synthetic Epistemology makes the same point from the epistemic side. Coherence is the criterion of validity, not correspondence; failure is recursive coherence loss, not deviation from truth. To validate *w* against an authoritative external annotation would be to assume the very substrate the framework denies and to import a static absolute into an architecture that eliminates static absolutes by design. This is the framework's central commitment, and the commitment is load-bearing throughout the corpus. The diagnostic of symbolic lock-in, the failure mode where a symbolic system becomes self-referentially closed and decoupled from the adaptive substrate it was built to track, depends on refusing the move that would make any symbolic system its own ground. A validation strategy that smuggles correspondence-truth back in would not test the framework. It would falsify the framework's ontology before any measurement was taken and then proceed as though that falsification were a methodological convenience. The resulting paper would be coherent within the conventions of empirical philosophy and incoherent within the framework it claimed to validate. That is precisely the drift my corpus names, a symbolic structure that has retained its surface form while losing contact with the persistence logic that made it meaningful.

The validation problem therefore cannot be solved by validation in the standard sense. It must be solved by something else, a structure of tests that operate inside the framework's commitments and impose discipline through the framework's own constraints rather than through appeals outside them. I propose that the question shifts. Not "does *w* match reality?", that question is unanswerable within the framework, and any answer to it would be drawn from outside this framework's ontology. The question instead becomes whether the operation that produces *w* coheres across the conditions the framework's ontology requires it to cohere across. That question is answerable, falsifiable, and load-bearing.

The seam from my prior induced-weights work generalizes upward when read in this register. Convergence across compressors that all share an inductive bias, such as all dictionary-based or all predictive-neural, is convergence within a philosophically homogeneous family. It demonstrates that the family's members agree with each other; it does not demonstrate that the structure being compressed exceeds the family's commitments. The dictionary class agreeing with itself is not evidence that the substrate is compressible by dictionaries; it is evidence that dictionaries cohere with one another, which we already knew. I argue that genuine intra-framework convergence requires estimator classes whose inductive biases disagree at the level of what counts as structure; classes

for which agreement, if it occurs, is informative about the substrate rather than about the family. The robustness claim must therefore be made across philosophical decoupling, not within philosophical agreement, and meeting that sharper requirement requires the explicit construction of an estimator class whose members are not variants of one compression philosophy.

Validation, in this register, becomes a coherence audit of this framework conducted in its own terms. The framework binds itself in advance to specific structural tests, names what each test would mean, and accepts the verdicts those tests produce. If the operation that produces w coheres across philosophically decoupled estimators, persists under perturbation in the way the framework's ontology requires, distinguishes structural intervention from interpretive intervention as the framework's central diagnostic demands, and remains the same operation across substrates of differing transparency, then the framework has earned operational status within the regime tested. If any of these conditions fail, the framework has correctly diagnosed where it actually applies, which is also a contribution. The resolution is not validation against an outside. This is what I will name meta-coherence, or the framework's measurement operation cohering with the framework's own commitments across conditions the framework specifies in advance, with each result configuration bound to a specific consequence before any data is touched.

Meta-coherence, named this way, is empty until specified. I propose the specification requires three distinct intra-framework signatures of self-consistent measurement. One tracking the convergence requirement, one tracking the persistence commitment, and one tracking the substrate-symbol distinction, together with a fourth admissibility gate, carried over from the prior admissibility commitments and made structurally explicit, that operates on the induced-w functional itself rather than on its within-domain behavior. Each of the four must be defined rigorously enough that an engineer reading them knows what to test. Each must be bound in advance to specific framework consequences. The next section does that work.

Three Intra-Framework Signatures of Self-Consistent Measurement, and the M5 Admissibility Gate

Meta-coherence specifies four conditions the induced-weights program must satisfy to register as operationally validated within the framework's own commitments. Three are signatures of self-consistent measurement, evaluated within a domain. The fourth is an admissibility condition on the weight functional itself, evaluated across domains. I take them in that order, then I argue why the architecture is a three-plus-one rather than a four.

2.1 R2, convergence under philosophical decoupling

The first signature targets the seam left open by prior induced-weights work. If induced w varies arbitrarily with the choice of compression estimator K, then w is a property of the compressor, not of the structure being compressed. The framework's universality claim is incompatible with that situation. It demands that w be invariant, modulo monotone transformation, across estimator classes whose inductive biases disagree at the level of what counts as structure.

Convergence is not an aesthetic preference. It is the operational signature of "structure that exceeds estimator." A claim that w tracks a real property of the substrate, rather than an artifact of the compression scheme through which the substrate is read, requires that estimators which would not otherwise agree be driven into agreement by the substrate itself. This recovers, at the meta-level, the discipline physics imposes when it requires independent measurement apparatuses with different physical principles to converge before a quantity is admitted as physical. A length measured by a ruler, a laser interferometer, and a radar return is admissible as a length because the three disagree about almost everything except the value they produce. R2 is that constraint translated into induced-weight space.

I name R2 as cross-philosophy convergence to mark the requirement explicitly; it is not enough that several compressors of one philosophical class agree. The dictionary class agreeing with itself is uninformative about the substrate. Convergence must hold across estimator philosophies that disagree at the level of inductive bias, before agreement carries weight as evidence.

2.2 R1, persistence prediction

Adaptive Realism defines reality through persistence across recursive coherence arbitration. If induced w names features as coherence-bearing without those features predicting persistence under perturbation, then w is not tracking what the framework says it tracks. The validation is internal, so the framework's own ontological commitment supplies the test, not an external truth-criterion.

I specify R1 as the requirement that high-w features participate demonstrably in the system's ability to persist. Under controlled perturbation, specifically perturbation of the substrate, not of the symbolic surface, features ranked high under induced w must show systematically distinct persistence behavior from features ranked low. The form of the persistence measure varies by domain (regime-inference recovery time in a synthetic stream, observed substitution tolerance in evolutionary data, post-event regime stability in institutional language), but the structure is constant;

w predicts persistence or it does not, and what it predicts is binary in form regardless of the domain's specific operationalization.

R1 is where the framework cashes its ontological commitment. If high-*w* features are not the features the substrate's persistence dynamics retain, the measurement operation has decoupled from the ontology that authorizes it. R1 failure is not a tuning problem; it is exposure.

2.3 R3, intervention asymmetry

The substrate-symbol distinction is load-bearing throughout the corpus. Symbolic lock-in names the failure mode in which symbolic structure persists despite decoupling from substrate dynamics; the symbol layer continues running on its own internal coherence after losing contact with what it was built to track. The framework's diagnostic of this failure depends on a distinction the framework's measurement operation must inherit. Structural change (modification of the substrate's constraint topology, of Ω, of the dynamics that generate the stream) is one kind of intervention; interpretive change (pure recoding, dynamics-preserving relabeling, symbolic substitution that leaves the substrate untouched) is another. The framework's claim is that these two are different in kind, not in degree.

R3 specifies the test. Under structural intervention, induced *w* must reorganize substantially. The substrate has changed, the relevance landscape has changed, so *w* must track the change. Under interpretive intervention, induced *w* must remain stable. The substrate has not changed, only the labels have, and *w* as a structural quantity must be invariant under relabeling that preserves dynamics. The test is a ratio between the magnitude of induced-*w* change under structural intervention divided by the magnitude under interpretive intervention. Asymmetry registers the substrate-symbol distinction; symmetry registers its collapse.

R3 is the most direct empirical test of the framework's central diagnostic. If *w* cannot distinguish substrate change from pure recoding, the framework's claim about symbolic lock-in is exposed at the level where the claim originates.

2.4 Joint necessity

R1, R2, and R3 are not additive options. Any singular one holding without the others is a structurally diagnostic failure, not a partial success. I specify the failure modes explicitly because the architecture depends on them.

R2 without R1, convergence without persistence prediction, registers a w that is internally consistent across estimator philosophies but ontologically empty. Several philosophically-distinct compressors agree on stable outputs; what they agree on does not predict what the framework's ontology says it should predict. The configuration names something stable but disconnected from persistence as the criterion of the real. The corpus already names this pattern where a symbolic system is stabilized in self-reference and decoupled from the substrate it claims to track.

R1 without R2, persistence prediction without convergence, registers a w that predicts persistence but is estimator-dependent. The chosen compressor produces a w that tracks the framework's ontology, but a different compressor would produce a different w. The universality claim is voided; what remains is conditional validity inside a chosen estimator philosophy, which the framework explicitly rejects as an admissible final state.

R3 without R1 and R2, asymmetry without convergence or persistence, registers a w that distinguishes intervention classes but does so unstably and without grounding. The measurement operation has inherited the substrate-symbol distinction at the level of perturbation classes it responds to differently, but it does not predict what the framework says it should predict and does not produce stable outputs across philosophical decoupling. This is a perturbation-class detector, not a coherence weight.

The three together form the minimum viable validation. Less is exposure at a specifiable layer.

2.5 The M5 admissibility gate

The R-signatures specified above are within-domain tests. Each evaluates whether the framework's measurement operation behaves coherently inside a single substrate-stream coupling. The framework's universality claim, however, is not within-domain, it is across-domain. The admissibility commitment named M5 (domain-translation invariance) in *Formal Foundation of Induced Coherence Weights* states the requirement directly: if a domain mapping φ preserves coherence up to monotone transformation, induced weights must be compatible under that mapping as well. M5 is not a fourth measurement signature. It is an admissibility condition on the induced-w functional itself, and it is distinct in kind from R1, R2, R3.

I operationalize M5 in the cross-domain context this paper specifies. The mapping φ does not operate at per-feature identity. Hidden Markov emission features, protein positions, and FOMC tokens have no natural one-to-one correspondence, and constructing one would import exactly the kind of domain-specific semantics M5 was designed to prohibit. φ operates at feature-class

semantics instead. Each domain admits a partition of its feature space into a coherence-bearing class and a noise class, by construction in D_1, by documented biology in D_2, and by documented institutional history in D_3. The partitions are committed at pre-registration and are not adjusted post-hoc.

Under matched estimator class K_i and matched ablation operator A_j, the induced-w functional applied in each domain produces a distribution over features. From that distribution, compute the ratio $r_d = \frac{\text{mean}_{w(\text{coherence–bearing class})}}{\text{mean}_{w(\text{noise class})}}$ for each domain $d \in \{D_1, D_2, D_3\}$. After rank-normalization of w within each domain, to strip cross-domain scale differences that have nothing to do with the functional's behavior, M5-admissibility requires that $r_{\{D_1\}}, r_{\{D_2\}}, r_{\{D_3\}}$ be consistent up to monotone transformation. The pre-registered threshold, $\frac{\max_d r_d}{\min_d r_d} \leq 2$ across all three domains, holding for at least three of the five estimator classes specified in Section 3.

The threshold is not arbitrary. A factor of 2 across rank-normalized ratios across three substrates of differing transparency, holding under at least three philosophically-decoupled estimators, is a strong invariance claim. A laxer threshold would admit functionals that drift across domains; a stricter one would demand more uniformity than the monotone-transformation clause permits. The 3-of-5 estimator requirement matches the convergence discipline at the within-domain layer; the same cross-philosophy logic that authorizes R2 authorizes the M5 evaluation, requiring philosophical agreement on the cross-domain consistency of the functional itself.

2.6 Why M5 is a gate, not an axis

If M5 fails, the induced-*w* functional is not domain-translation invariant. The framework's universality claim is foreclosed at the functional level, regardless of how the within-domain R-signatures resolve. R1, R2, and R3 may all pass within each domain individually while M5 fails. That configuration is informative, but it is not the framework's universality claim. It would mean the framework produces a coherent measurement operation in each domain while producing a *different* operation in each domain. The framework does not claim that. It claims one operation, applicable universally, and M5 tests the claim at the level the claim actually lives.

The reverse configuration, M5 passing while R-signatures fail within domains, is also informative. The induced-*w* functional would be consistent across domains in what kind of structural distinction it tracks, while the within-domain measurement operation fails its own internal tests. This configuration is structurally distinct from M5 failure and is bound to a separate interpretation in the

outcome matrix. It indicates that the functional is universal in its target but not yet operational in its execution; the same kind of measurement, performed badly in each instance.

The gate logic follows from the architecture. M5 is evaluated first; its result conditions the interpretation of the R-signature outcome matrix. M5-pass licenses the eight-cell matrix at full strength. M5-fail does not void the matrix; it conditions its reading. The matrix is computed and reported in either case, but the universality claim is foreclosed under M5-fail regardless of how the cells resolve.

2.7 The admissibility-measurement separation

The cleanest read of the architecture is that R1, R2, and R3 ask whether the measurement operation works inside a domain. M5 asks whether the functional is the same functional across domains. Both must hold for the framework's universality claim to register as supported. Either failing is exposure, but at different layers. R-signature failure exposes measurement, M5 failure exposes the functional itself, and each layer admits distinct revision pathways for the corpus.

The four conditions, taken together, are what meta-coherence specifies. Three within-domain signatures of self-consistent measurement plus one across-domain admissibility gate on the functional. None is reducible to the others. Each must be specified to the point of falsifiability, and each is.

The specification, however, is empty until grounded. Meta-coherence demands a meaningful estimator class capable of philosophical decoupling, without which R2 has no traction and M5's matched-estimator-class requirement has no instances. It demands operational definitions of persistence per domain, without which R1 cannot be computed. It demands a principled distinction between structural and interpretive interventions in concrete cases, without which R3 reduces to an a priori category exercise. Finally, it demands explicit feature-class partitions per domain, without which M5's ratio computation has no operands. The next section begins with the first.

The Estimator and Operator Classes

The R-signatures and the M5 admissibility gate are empty until grounded in concrete estimator and operator specifications. Section 2 named what cross-philosophy convergence requires, an estimator classes whose inductive biases disagree at the level of what counts as structure, such that agreement among them is informative about the substrate rather than about the family. This section commits to a specific choice of five estimators and three ablation operators, justifies each on

philosophical-decoupling grounds, and specifies the 5×3 grid as the shared infrastructure that R2 and M5 both read against different axes.

3.1 Estimator class K

The robustness criterion inherited from prior induced-weights work requires that the estimator class span genuinely different inductive biases, not variations of a single compression philosophy. I specify five estimators.

K_1 is dictionary compression, instantiated through zstd at fixed compression level. Its inductive bias is repetition-based regularity capture. The substrate is read as a stream of recurring sub-sequences, and structure is whatever the dictionary's pattern-matching apparatus can encode efficiently. Long-range structure invisible at the dictionary's window is invisible to K_1.

K_2 is arithmetic-coded n-gram minimum description length, averaged across $n = 3, 4, 5$ with model-cost paid explicitly under MDL accounting. Its inductive bias is explicit statistical modeling at fixed conditional-context length. The substrate is read as a Markov-of-order-n source, and structure is whatever the conditional distributions encode. Coalitional structure spanning more than n positions is invisible to K_2; structure encoded in the joint statistics of those n positions is exactly what K_2 tracks.

K_3 is small-transformer negative log-likelihood under prequential coding. The transformer is trained online over the stream with predictions logged before each update; the cumulative prequential loss serves as the description length. The inductive bias is predictive-neural. Structure is whatever a parameterized predictor with attention-based context aggregation can learn from the stream's history. K_3's context window is broad and its parameterization is non-linear; long-range dependencies and non-Markovian structure are within reach where K_1 and K_2 cannot see them.

K_4 is hidden Markov model and state-space-model minimum description length under model-class search. The MDL search ranges over hidden-state cardinality and emission-distribution structure; the selected model's description length plus the observation likelihood gives the coherence proxy. The inductive bias is generative state-space. The substrate is read as a latent-state process emitting observations, and structure is whatever the latent dynamics encode. Long-range temporal coupling mediated through state continuity is within reach.

K_5 is Lempel-Ziv complexity proper, in the Kaspar-Schreiber sense, distinct from the LZ77 compression scheme that informs K_1. K_5 counts the number of distinct factors required to reproduce the sequence under a specific parsing rule; it is a measure of algorithmic structural

complexity, not a coding scheme that produces an encoded output. The inductive bias is algorithmic. Structure is whatever the parsing rule registers as factor diversity, with no explicit model and no compression artifact intermediating between the sequence and the complexity measure.

The fifth estimator is critical, and the reason is structural rather than operational. K_1 through K_4 are coding schemes. Each produces an encoded representation whose length is the description length, and each carries an implicit ontology of "good code." Agreement among K_1–K_4 on what constitutes coherence-bearing structure is consistent with the possibility that all four are tracking codeable structure rather than substrate structure, and that the framework's apparatus, run through any of them, would converge on the same artifact of "what compresses well." K_5 does not produce a code. It registers structural complexity through factor enumeration without intermediating through an encoded representation. If induced-w convergence holds across K_1–K_5, the substrate is responding to something that exceeds “good codes” and agreement that survives the philosophical disagreement between coding and non-coding registers as cross-philosophy convergence in the sharpest sense available. Convergence holding only across K_1–K_4 would register as family-internal consistency. Informative, but not sufficient for R2 at full strength.

3.2 Ablation operator class A

The estimator-class problem nests inside an operator-class problem. The induction pipeline computes a relevance score $\rho(x)$ by ablating x from the stream and measuring the induced change in the coherence proxy; the form of the ablation determines what kind of contribution ρ tracks. A single ablation operator imports a single inductive bias about how features contribute to coherence, and convergence across estimators under one operator does not test convergence across notions of contribution. I specify three operators.

A_1 is leave-one-out. Each feature is ablated independently; $\rho(x)$ is the marginal coherence delta. The inductive bias is independent contribution; a feature's relevance is the change in coherence when only that feature is removed, holding everything else fixed. Coalitional structure, coherence contributions that emerge only from joint feature presence, is systematically underweighted by A_1, because LOO computes the marginal effect with the coalition partner still present, capturing only the share of joint contribution attributable to the single feature in isolation.

A_2 is Shapley-style coalitional ablation, evaluated over $k = 64$ sampled coalitions per feature. Shapley values average a feature's marginal contribution across many ablation contexts, with coalitions sampled from the full feature-set powerset under weights that satisfy the Shapley axioms.

The inductive bias is interactional; a feature's relevance is its expected marginal contribution averaged across all subsets, capturing joint contributions that LOO misses. The sampling depth is fixed in advance to control computational cost while providing unbiased estimation under the Shapley sampling theorem.

A_3 is group ablation by correlation cluster. Features are clustered by Pearson correlation at threshold $|r| > 0.5$; ablation removes entire clusters jointly; $\rho(x)$ for a feature is its share of the cluster's coherence contribution. The inductive bias is redundancy-aware; a feature's relevance is computed against the structure of correlations it participates in, with redundant features sharing rather than competing for relevance. A_3 captures structure that A_1 over-attributes to individual features (when the contribution is actually shared across a correlation cluster) and that A_2 partially recovers but at higher computational cost and with different mathematical commitments.

The three operators span the philosophical decoupling at the operator axis. A_1 holds individual contribution as primary; A_2 holds interactional contribution as primary; A_3 holds redundancy structure as primary. Without operator-class diversity, the framework's measurement operation imports whichever bias the chosen operator carries, and the resulting w reflects that bias as much as it reflects the substrate. Recursive coherence theory should be least biased toward independent contributions; the framework's ontology emphasizes recursive reinforcement, which is interactional by structure. A measurement operation that defaults to LOO and stops there is a measurement operation tilted away from the kind of structure the framework most claims to track.

3.3 The 5×3 grid as the primary axis

The product of estimator class K and operator class A yields fifteen (K, A) cells per domain. Each cell produces an induced-w distribution over the domain's features. The grid is the primary axis of the empirical program.

R2 evaluation is pairwise across the grid within a domain. For each pair of cells (K_i, A_j) and (K_k, A_l), Spearman rank correlation between the two induced-w distributions is computed. The cross-philosophy median Spearman, evaluated across the cell pairs that span philosophically-decoupled estimator classes (and, separately, philosophically-decoupled operator classes), supplies the load-bearing R2 signal. Within-philosophy block correlations, among $K_1/K_2/K_3$ within the coding-scheme class, or among A_1/A_2 within the per-feature class, are reported separately as a discriminative diagnostic. The contrast between cross-philosophy and within-philosophy correlations is itself

informative; high within-philosophy correlation with low cross-philosophy correlation would indicate exactly the family-internal-consistency failure mode Section 2 named.

The fifteen cells are not differently weighted. Each contributes one induced-w distribution, and the grid's structure, five estimator philosophies crossed with three operator philosophies, is what the convergence claim is evaluated against. The grid runs identically in every domain.

3.4 The coherence proxy

Inside the induction pipeline, each estimator computes a coherence proxy $\hat{C}$ from which relevance $\rho(x)$ is derived under the chosen operator. The proxy admits two principal forms in prior engineering work: a compression-delta form (proxy = baseline description length minus description length conditional on retaining x) and a predictive-logloss form (proxy = predictive log-likelihood, with relevance derived from predictive contribution). I fix the predictive-logloss form as the load-bearing form across all domains and all (K, A) cells, with the compression-delta form run as a secondary sanity check on D_1 only.

The reason is architectural. The load-bearing empirical claim is $K \times A$ robustness, that induced w converges across philosophically-decoupled estimators and operators, and that the cross-domain functional is admissible under M5. Letting the proxy-family vary as a third primary axis would multiply the cells without adding philosophical decoupling. Predictive-logloss and compression-delta are two routes through the same induction pipeline, not two competing inductive biases about what counts as structure. Treating proxy-family as a sanity check rather than a primary axis preserves the decoupling load on K and A, where it belongs, and registers any proxy-family-dependent drift as a diagnostic finding in D_1 rather than as a contamination of the primary results.

3.5 The grid's role in M5

The same 5×3 grid is run identically across D_1, D_2, and D_3. R2 reads the grid one way; M5 reads it another.

R2 reads horizontally within each domain. For each domain, the fifteen induced-w distributions are compared pairwise; the cross-philosophy convergence claim is evaluated against the within-domain pattern of agreements and disagreements. R2 is a within-domain test of measurement convergence across estimator and operator philosophies.

M5 reads vertically across domains under matched estimator and operator philosophy. For each (K_i, A_j) cell, the induced-w distribution in D_1 is compared to the induced-w distribution in D_2 and to

the induced-w distribution in D_3; the cross-domain ratio of mean-w over coherence-bearing features to mean-w over noise features is computed per domain and the consistency of those ratios across domains is evaluated under the threshold specified in Section 2.5. M5 is an across-domain test of functional consistency at fixed philosophical commitment.

The two tests are mutually informative but not redundant. R2 evaluates whether different philosophies agree within a domain about what counts as coherence-bearing. M5 evaluates whether the same philosophy produces a consistent kind of measurement across domains. The grid is the shared infrastructure; the two tests read different cuts of it, and the architecture of meta-coherence requires both.

The estimator class is now specified, the operator class is now specified, the grid is now committed, and the proxy-family choice is locked. The program needs domains across which to run the grid. The next section establishes what "domain" means in a framework that forbids correspondence truth, and how three carefully chosen test surfaces carry both the within-domain measurement load and the across-domain admissibility load.

The Transparency Hierarchy

The empirical program specified in Sections 2 and 3 commits to three domains. Before the domains themselves are specified, the conceptual frame governing why these three and not others must be established, because the natural reading of "test the framework on synthetic data and real data" imports an assumption the framework cannot accept. The assumption is that real data touches the world while synthetic data does not, and that empirical streams have an epistemic privilege over generated streams in virtue of their contact with an unmediated substrate. This reading depends on correspondence-truth. Synthetic Epistemology refuses correspondence-truth. The reading therefore does not survive contact with the framework whose validation it would govern, and a different conceptual frame is required.

4.1 The synthetic-empirical illusion

The standard hierarchy treats "synthetic" as a weak class of data, useful for debugging, instructive for sanity checks, but unable to validate claims about the world because the world is not what generated it. It treats "empirical" as a strong class, because the empirical stream came from the substrate the framework purports to describe. The hierarchy is intuitive and, within correspondence-truth epistemology, correct. Within Synthetic Epistemology it dissolves.

All measurement is compression. The Lorenz attractor we integrate numerically is a substrate; its trajectory is generated by a fully specified dynamical system, and what we record is a compression of that trajectory through whatever instrument we use to record it.

The protein sequencer's output is a substrate; the trajectory is generated by molecular evolution operating on physical chemistry, and what we record is a compression of that trajectory through alignment algorithms, redundancy filters, and curation conventions. The FOMC transcript is a substrate; the trajectory is generated by institutional decision-making operating on economic conditions, and what we record is a compression of that trajectory through committee deliberation, drafting protocols, and publication norms. None of these streams is unmediated. None has direct contact with a bedrock the others lack. They differ along a different axis, what generated them and how transparent the generator is to us. That axis is the one this framework uses.

The synthetic-empirical illusion is the residue of correspondence-truth thinking inside an epistemology that has already abandoned it. Replacing the illusion is not optional. The validation program cannot be coherent with the corpus it extends if the program imports the very ontology the corpus eliminates.

4.2 The transparency hierarchy

I propose three domains that span a substrate-stream transparency gradient. The gradient is defined by how transparently the generator is accessible to specification, not by how much epistemic privilege the resulting stream possesses.

D_1 is the full-transparency case. The generator dynamics are fully specified, a hidden semi-Markov source with structured emissions, with every parameter, every transition, every coupling, every noise class committed in advance. The substrate is fully transparent because it was written. The stream is what the writer's instruments recover from observation, with no privileged access to the specification.

D_2 is the structural-transparency case. The generator dynamics are partially formalized. Molecular evolution and protein folding are understood at intermediate resolution; selection pressures are real, conservation patterns are documented, structural constraints are characterized through decades of experimental and computational work. The substrate is not opaque, but it is not fully specified either. The generator is partially knowable, partially inferred.

D_3 is the opaque-substrate case. The generator, institutional decision-making under economic constraints, operating through committee dynamics, political pressures, and the deliberate

ambiguity that characterizes central-bank communication, is inferable only through the stream itself and through whatever historiographic apparatus the analyst chooses to admit. The substrate is not accessible to specification in the way D_1 and D_2 are. What is observed is the symbolic surface, and inferences about the generator are themselves compressions running on that surface.

This is not a quality ranking. The three domains are not arranged from "weak" to "strong" or from "controlled" to "real." They are arranged along a single structural axis, how much we know about what generated the stream, and the framework's universality claim makes a specific prediction along that axis. If the three R-signatures hold across the full gradient, and M5-admissibility holds across the gradient, the framework's measurement operation and the induced-w functional are jointly substrate-transparency-invariant. That is the strongest internal-consistency claim available under Synthetic Epistemology, and it is what universality, in the framework's own terms, would mean. If signatures hold for D_1 but degrade for D_2 and D_3, the framework requires substrate transparency in order to validate, which sharply restricts its universality to controlled or partially-controlled regimes. If signatures hold for D_2 and D_3 but not D_1, the synthetic generator is exposing a flaw in the inference machinery that natural streams happen to mask; a coherent and informative finding, because it would mean the framework's apparatus is sensitive to substrate features the natural-stream tests cannot diagnose. Each pattern is a contribution; each is bound to a specific framework consequence in Section 8.

4.3 Why D_1 is not circular

The standard objection to D_1 runs as follows: "you built the stream to be coherence-bearing, you find that it is coherence-bearing, what did you prove?" The objection depends on a misreading of what D_1 tests.

D_1 does not test whether the framework correctly models reality. It tests whether the measurement operation, given only the symbol stream, recovers structural properties that are present in the generator but invisible at the surface. The measurement operation has no privileged access to the generator's specification. It sees what every other estimator sees, a sequence of symbols. The generator's structure, the regime layer, the long-range coupling, the coalitional encoding, the drift carrier, is recoverable from the surface only if the inference machinery is coherent with the substrate that produced it. If the machinery is coherent, recovery happens. If it is not, recovery fails.

This is exactly the move physics makes with controlled lab systems. The double-pendulum equations are written down by the experimenter; the experimenter then asks whether instruments measuring

only position and time can recover the chaotic dynamics the equations specify. Recovery does not validate the equations against reality; the equations are stipulated, not tested. Recovery validates the instruments. If the instruments cannot recover what the experimenter built, the instruments are inadequate to the substrate. If they can, the instruments are coherent with substrates of that structural class.

D_1's asset is substrate-stream coupling transparency. Because the generator is specified, the test has resolution; we can ask not just whether the measurement operation works but exactly which structural properties it recovers and which it misses, under which estimator-operator combinations. That granularity is unavailable in opaque-substrate domains, where the question "did the measurement operation recover what was there?" cannot be sharpened beyond "did it produce results that look right." The transparency is what makes diagnostic precision possible. Refusing D_1 on circularity grounds amounts to refusing the resolution it provides.

4.4 Why D_2 does not require Pfam as ground truth

D_2 uses protein domain families. The natural move would be to treat Pfam annotations as ground truth; induced w is correct if and only if it agrees with the conserved positions Pfam identifies. This move would make Pfam the adjudicator and induced w the candidate. This framework cannot accept that arrangement.

Pfam annotations are themselves a compression. They are the accumulated product of biology's prevailing inductive biases, run by a community of researchers performing their own coherence-induction over evolutionary data, with conservation models, structural homology arguments, and curation conventions doing the work that estimators do in the framework's apparatus. Treating Pfam as ground truth would mean induced w is correct iff it agrees with biology's prevailing compression of the same data. That is deference to incumbent symbolic authority, and the framework names that exact pattern as a failure mode. The corpus's diagnostic of symbolic lock-in includes the institutional case where a community's accumulated compression hardens into the authority against which all subsequent compressions are judged.

Pfam enters D_2 instead as one more compression among several. The framework's induced-w functional is run through the same pipeline that produces every other w in the program. Pfam's conservation annotations are run through a parallel functional that takes Pfam's curation outputs as the relevance score. The two are then compared. Convergence is one more meta-coherence signal. Biology's prevailing compression and the framework's compression agree on what the substrate

exposes, but Pfam does not arbitrate. Cells where induced w diverges from Pfam are reported as informative, not as failures of induced w. The divergence may indicate that the framework's apparatus is detecting structure Pfam's curation conventions miss, or that Pfam is detecting structure the framework's apparatus misses, or that the two are tracking partially different aspects of the same evolutionary substrate. Each reading is a finding. None positions Pfam as judge.

4.5 Why D_3 is the sharpest test

Institutional language is where the framework's diagnostic of symbolic lock-in lives, and it is where that diagnostic faces its most direct empirical challenge. The Federal Open Market Committee produces statements that operate at two levels simultaneously, the substrate of monetary policy (rate decisions, balance-sheet operations, regime shifts) and the symbolic surface of forward guidance (the deliberate language of expectations management, framed in conventions that evolve independently of underlying policy in significant ways). The framework's claim is that induced w can distinguish these two, and that structural change at the substrate produces a different kind of weight reorganization than pure linguistic-convention shift at the surface.

If induced w cannot make this distinction in FOMC data, the framework's central diagnostic collapses at exactly the empirical point where it is most consequential. The corpus's claim about symbolic lock-in, institutional, governmental, scientific, and cultural, depends on the substrate-symbol distinction being real and detectable through the framework's apparatus. D_3 tests whether it is. The risk is real, and the reward is proportional. R3 in D_3 is the load-bearing test for the framework's institutional-dynamics claim, and it earns or fails to earn that claim through specific, documented breakpoints; structural ones the historiography records and linguistic ones the historiography also records, with the two distinguishable in advance. The pre-registration commits to that distinction before any analysis runs. The framework either inherits the distinction or exposes its absence.

4.6 The hierarchy as test surface

The gradient is predictive. The framework's universality claim, translated into Synthetic Epistemology's terms, is operationally equivalent to substrate-transparency-invariance of the three R-signatures plus M5-admissibility of the functional itself. The transparency hierarchy converts a vague universality claim into a specific empirical pattern that holds or does not hold across a structural axis. Either the pattern holds across the full gradient, or where it breaks down

characterizes the framework's actual scope. Both outcomes are contributions, and each is bound to a specific revision pathway for the corpus in advance.

The hierarchy also fixes a methodological point that would otherwise float. The synthetic-empirical illusion would treat results in D_1 as preliminary and results in D_3 as final, with D_2 as a middle case to be weighed by the analyst. The transparency-gradient framing eliminates that asymmetry. All three are formally synthetic in the framework's sense, all three are streams produced by generators whose mediation is acknowledged, and the framework's universality claim binds them with equal weight. D_1 supplies diagnostic resolution; D_2 supplies evolutionary substrate; D_3 supplies institutional opacity. The three together test what the framework claims, at the level the claim actually lives.

4.7 Domain partition for M5

The hierarchy fixes the feature-class partitions M5 requires. M5 admissibility, as specified in Section 2.5, computes the ratio of mean induced w over coherence-bearing features to mean induced w over noise features within each domain, then tests whether the ratio is consistent across domains under matched estimator class. The partition between coherence-bearing and noise classes is therefore not optional and cannot be derived from induced w itself; that would be circular at the partition level. The transparency gradient supplies the partition logic without circularity. In D_1, the partition is by construction. The generator's specification names which features carry coherence-bearing structure (properties A, B, C, D, in the four-property design) and which features are noise (n_1, n_2, n_3, the distractor classes). The partition is committed at pre-registration alongside the generator parameters and is not adjustable post-hoc. In D_2, the partition is by documented biology. Catalytic residues, fold-defining positions, and conserved cores are the coherence-bearing class; surface-exposed non-functional residues and bottom-conservation-quintile positions are the noise class. The partition is committed per family at pre-registration, drawn from the existing biological literature, and locked. In D_3, the partition is by documented institutional history. Tokens flagged in published Fed-communication scholarship as substrate-policy-bearing constitute the coherence-bearing class; conventional discourse markers, procedural language, and committee-membership references constitute the noise class. The partition is a curated lexicon, drawn from existing institutional-economics literature, published with the pre-registration, and locked.

Each partition operates within its domain's transparency tier. D_1's partition is exact because the substrate is written; D_2's partition is constrained by documented biology; D_3's partition is constrained by documented institutional history. None depends on induced w itself, which is what

M5's gate logic requires. The partitions feed directly into M5 evaluation in Section 8 and are referenced in the domain-specification sections that follow.

The conceptual frame is now established. The transparency gradient governs which domains the program tests, the partition logic governs how M5 is evaluated, and the joint substrate-transparency-invariance claim names what universality would mean if it held. Each domain still requires a concrete generator specification, a persistence operationalization, an intervention-class definition, and the feature-class partition just exposed, before any of this becomes runnable. Section 5 specifies D_1 in full detail; Sections 6 and 7 do D_2 and D_3 respectively.

Domain D_1: Hidden Semi-Markov Source with Structured Emissions

D_1 is the full-transparency case of the hierarchy. Its purpose in the program is diagnostic: because the substrate is written, the test of the measurement operation has resolution unavailable in the other domains. This section specifies the generator architecture, the emission features, the four-property diagnostic design, the parameter calibration protocol, the operationalizations of R1, R2, and R3 in this domain, the feature-class partition that grounds M5, and the property-recovery cross-tabulation that constitutes the section's empirical claim.

5.1 Generator architecture

The generator is a hidden semi-Markov regime layer Z_t over three states, with structured discrete emissions X_t conditioned on Z_t and on a small set of auxiliary indices internal to the generator. Hidden semi-Markov rather than hidden Markov is the load-bearing structural choice; explicit duration distributions decouple regime sojourn from the geometric-tail constraint that hidden Markov dynamics would impose, and regime-shift detection is exactly the kind of structural feature whose recovery should not be confounded by an artifact of the emission model.

Sojourn durations are drawn from negative binomial distributions with mean dwell ≈ 200 steps and dispersion parameter chosen to give moderate variance (coefficient of variation ≈ 0.4). The mean is calibrated against the long-range-coupling lag τ specified below. Regimes must persist long enough that the long-range structure inside a regime is learnable, but not so long that the stream becomes effectively single-regime over feasible analysis windows. The dispersion is calibrated to avoid both deterministic-duration trivialization and exponential-tail contamination.

The regime transition matrix is row-stochastic over three states with sparsity controlled by a fixed off-diagonal mass parameter; transitions to either non-current state are equiprobable. The matrix is

irreducible and aperiodic, ensuring ergodicity and a well-defined stationary distribution. Regime structure must be present and recoverable, but not trivially dominant. The emission distributions are calibrated such that no single feature's marginal entropy distinguishes regimes at decisive strength under naive frequency analysis. Regime recovery requires the measurement operation to integrate information across features and across time, which is what the structural design demands.

The stream length per simulated dataset is $N = 50{,}000$ steps, which yields approximately 250 regime sojourns on average. Replication is by independent random seed; the pre-registered analysis runs on $M = 20$ replicate streams, with all reported statistics aggregated across replicates.

5.2 Emission feature specification

The observation at each step is $X_t = (f_A, f_B, f_{C_1}, f_{C_2}, f_D, n_1, n_2, n_3)$; eight discrete features, each over an alphabet of size 8, encoded as a length-8 categorical vector.

f_A is a regime indicator. Its emission distribution is regime-conditional. Under each Z_t state, f_A draws from a distinct categorical distribution over its alphabet, with the three regime-conditional distributions chosen to have moderate but not maximal pairwise total-variation distance. f_A is the cleanest single-feature signal of regime state; its presence in the design tests whether the measurement operation can recover the most accessible coherence-bearing structure.

f_B encodes long-range temporal coupling. Its value at step t is a deterministic function g of X at step $t - \tau$, with $\tau = 50$; $f_{B(t)} = g(X_{\{t-\tau\}})$. The function g is a fixed pseudorandom mapping from the eight-feature observation space into the f_B alphabet, committed at pre-registration. The coupling becomes detectable only when the measurement operation's effective context window L exceeds 100 steps; estimators with shorter context cannot recover f_B's structural role regardless of operator choice. f_B in the design tests whether the measurement operation can recover long-range temporal structure that local statistics cannot expose.

f_{C_1} and f_{C_2} form a coalitional pair. Each individually emits according to a high-entropy categorical distribution that is approximately marginally uniform; individually, neither feature's marginal statistics distinguish regime state above noise. Jointly, however, the pair encodes a pseudorandom function of the regime state Z_t and a slow-time index $s_t \in \{1, 2, \ldots, 4\}$ that cycles deterministically with period equal to the mean regime sojourn. The pair is constructed such that $(f_{C_1(t)}, f_{C_2(t)})$ determines (Z_t, s_t) up to an ambiguity that resolves over windows of length comparable to the sojourn duration. f_{C_1} and f_{C_2} in the design test whether the operator class can recover coalitional structure that no leave-one-out analysis can detect.

f_D is a drift carrier. It emits according to a categorical distribution whose parameters drift slowly across the stream, with drift rate calibrated such that drift accumulates significantly over a regime sojourn but remains within-regime-detectable. The drift correlates with regime transition probability; as drift accumulates, the conditional probability of transition rises, with the relationship encoded as a fixed monotone mapping committed at pre-registration. f_D in the design tests whether high induced w on f_D's drift component predicts persistence, which is the R1 signal in this domain.

n_1, n_2, and n_3 are distractor noise classes. n_1 emits uniform high-entropy noise; marginally, its distribution is uniform over its alphabet, and its values are independent across time and across other features. n_2 emits Zipf-structured noise; its marginal distribution is power-law, giving the surface appearance of structure that frequency-based estimators may mistake for coherence-bearing. n_3 emits low-frequency noise; a small subset of alphabet symbols dominates with high probability, again giving surface statistical structure without coherence-bearing role. The three distractor classes test whether the measurement operation distinguishes coherence-bearing structure from frequency-dominant noise, the central failure mode the engineering interface paper flagged for benchmark design.

5.3 The four properties as diagnostic axes

The emission design encodes four structural properties, each mapped to a specific measurement-operation capability the property tests.

Property A (regime-indicator recoverability) tests whether induced w concentrates appropriately on f_A under estimator classes with even modest context. The recoverability is LOO-tractable; f_A's marginal contribution to the coherence proxy is direct, and any reasonable estimator paired with any reasonable operator should produce induced w on f_A that exceeds induced w on the distractor classes.

Property B (long-range temporal coupling) tests whether the measurement operation recovers structure encoded at lag $\tau = 50$. Recovery is estimator-dependent. State-space estimators (K_4) and predictive-neural estimators with adequate context (K_3) should recover f_B's role; dictionary compression with short windows (K_1) and low-order n-gram MDL (K_2 at $n = 3$) should miss it. The pattern is informative; which estimators recover Property B characterizes the inductive-bias structure of the estimator class.

Property C (coalitional structure) tests whether the operator class recovers structure that emerges only from joint feature presence. Recovery is operator-dependent. A_1 (LOO) systematically misses

Property C: ablating f_{C_1} alone, with f_{C_2} still present, leaves the joint encoding partially intact and produces a marginal coherence delta far below the joint coalition's contribution. A_2 (Shapley) recovers Property C when the sampling depth $k = 64$ is sufficient to capture the f_{C_1}/f_{C_2} coalition with adequate probability. A_3 (group ablation) recovers Property C when the f_{C_1}/f_{C_2} correlation is detected by the clustering threshold and the pair is ablated jointly.

Property D (drift-persistence coupling) tests whether high induced w on f_D predicts persistence under perturbation, which is the R1 signal in D_1. Recovery requires that the measurement operation register f_D's slow drift component as coherence-bearing; a non-trivial requirement, because f_D's marginal entropy is comparable to the distractor classes, and only its temporal-trajectory structure (its slowness, its trend) distinguishes it.

5.4 Parameter calibration protocol

The generator's parameters (sojourn distribution parameters, B-coupling lag τ, C-pair encoding strength, D-drift rate, alphabet sizes, transition matrix off-diagonal mass) are calibrated before pre-registration through a small grid search. The objective of the search is structural; lock parameters such that no single property explains more than 40% of total $\hat{C}$ variance under the K_3 baseline estimator with A_1 operator (chosen as the calibration reference because it is the most generic combination), and such that all four properties are in principle recoverable above the noise floor that $n_1/n_2/n_3$ establish.

The 40% ceiling prevents Property A from dominating the diagnostic signal; without it, the cross-tabulation in 5.7 would be uninformative because Property A would saturate the induced-w distribution under any halfway competent measurement operation, masking the more discriminative Properties B and C. The recoverability floor prevents the design from making any property invisible to even the best estimator-operator combination, which would degrade the cross-tabulation symmetrically.

Calibration is a procedural freedom. Pre-registration locks the parameters that calibration produces, not the calibration procedure itself. Once parameters are locked, the analysis commits to those values, and any post-hoc parameter adjustment is excluded by the pre-registration discipline.

5.5 Operationalization of R1, R2, R3 in D_1

R2 (cross-philosophy convergence) is computed pairwise across the 5×3 grid as specified in Section 3.3. For each pair of cells, Spearman rank correlation between the induced-w distributions over the

eight features is computed, with bootstrap confidence intervals constructed by time-window resampling at block length equal to one mean sojourn duration. The cross-philosophy median Spearman, evaluated over cell pairs spanning philosophically-decoupled estimator classes, supplies the R2 signal. The R2 pass threshold is the global threshold from Section 8.

R1 (persistence prediction) operationalizes through a perturbation protocol. For each feature class, coherence-bearing and noise, controlled perturbation is introduced. Each feature's emission distribution is corrupted by a small mixing perturbation (a fixed-magnitude convex combination with a uniform distribution), applied at a designated time point in a perturbation-replicate stream. Two persistence measures are computed. The first is regime-inference recovery time τ_{rec}: the number of steps after the perturbation onset before a Viterbi decoder applied to the perturbed stream recovers regime estimates within a calibrated agreement threshold of the unperturbed-stream Viterbi estimates. The second is predictive horizon h_{pred}: the number of steps after perturbation over which the perturbed stream's predictive log-likelihood under the K_3 baseline estimator remains within a calibrated band of the unperturbed stream's. R1 tests whether high-w features yield systematically worse persistence under perturbation than low-w features. That is, whether perturbing what the measurement operation says is structurally important produces a larger persistence cost than perturbing what it says is not.

R3 (intervention asymmetry) operationalizes through paired structural and interpretive interventions. Structural intervention: modify the regime transition matrix (a fixed perturbation of off-diagonal mass) or modify f_D's drift rate (a fixed perturbation of the slow-component parameters); rerun the induction pipeline; compute the induced-w distribution under the perturbed substrate; measure the magnitude of induced-w change as $\|\Delta w\|$ in L1. Interpretive intervention: apply a deterministic permutation to one feature's emission alphabet, chosen to preserve all probability structure (the permuted distribution is identical to the original under relabeling); rerun the induction pipeline; compute $\|\Delta w\|$. The R3 ratio is the structural $\|\Delta w\|$ divided by the interpretive $\|\Delta w\|$, computed per (K, A) cell and aggregated across the grid. The R3 pass threshold is the global threshold from Section 8.

5.6 D_1 feature-class partition for M5

The feature-class partition for M5 evaluation in D_1 is by construction; the coherence-bearing class is $\{f_A, f_B, f_{C_1}, f_{C_2}, f_D\}$, and the noise class is $\{n_1, n_2, n_3\}$. The partition is exact. It is not derived from induced w or from any analysis of the simulated streams; it is a property of the generator's

specification. M5 evaluation, as specified in Section 2.5, computes mean induced *w* over the coherence-bearing class divided by mean induced *w* over the noise class, per (K, A) cell, and feeds this ratio into the cross-domain consistency check at the M5 gate in Section 8.

D_1's partition is the cleanest of the three domains because the substrate is written. D_2's partition is constrained by documented biology; D_3's is constrained by documented institutional history. The transparency gradient maps directly onto the partition's epistemic status; D_1's partition is exact, D_2's is documented, D_3's is curated.

5.7 Property recovery as the section's diagnostic deliverable

The section's empirical claim is the cross-tabulation of property recovery across the 5×3 grid. For each (K, A) cell, the analysis reports which properties (A, B, C, D) the cell's induced-*w* distribution recovers, where recovery is defined as the property's signal features ranking systematically above the distractor classes by a pre-registered margin. The cross-tabulation produces a 15-row, 4-column table whose pattern is the diagnostic deliverable.

The framework does not "win" or "lose" D_1 at the level of a single threshold. It is characterized by which estimator-operator combinations recover which properties. $K_1 \times A_1$ should recover A and miss B, C, D. $K_3 \times A_2$ should recover A, B, C, and partially recover D. $K_5 \times A_3$ should recover A and C and supply a non-coding-scheme cross-check on B and D. Each cell's recovery profile is a fingerprint of its inductive bias under the generator's specification, and the aggregate pattern characterizes the resolution and blind spots of the measurement operation as such; not as an evaluation of whether the framework succeeds, but as an evaluation of where the framework's apparatus has traction and where it does not.

D_1's transparency is its asset, both for the diagnostic granularity of the cross-tabulation and for grounding M5's feature-class partition without ambiguity. D_2 trades full transparency for richer evolutionary dynamics. The next section specifies how the framework handles a domain where the generator is partially knowable, partially opaque.

Domain D_2: Pfam Protein Domain Families

D_2 is the structural-transparency case. The generator, molecular evolution operating on physical chemistry under selection, is partially formalized. Selection pressures are real and quantifiable, conservation patterns are documented across decades of comparative genomics, and structural constraints on protein folding are characterized at intermediate but not exhaustive resolution. The

domain trades D_1's exact transparency for richer evolutionary dynamics that no synthetic generator could produce at comparable scale. This section specifies family selection, stream construction, persistence operationalization, Pfam's role as comparison rather than adjudicator, intervention protocols, and the M5 feature-class partition.

6.1 Family selection

The analysis runs on 20–30 Pfam families, with the exact count fixed at pre-registration and no post-hoc adjustment. Selection is stratified across four axes simultaneously, none of which is allowed to dominate.

Ortholog depth is the first stratification axis. Each selected family must contain at least 500 sequences after redundancy filtering at 90% sequence identity. The threshold ensures that observed-substitution-rate computations have adequate statistical power per position; below it, position-wise tolerance estimates become noise-dominated, and the persistence operationalization in 6.3 loses traction. Families with 500–1500 orthologs and families with 1500+ orthologs are sampled in approximately equal proportion to avoid selection bias toward either over-curated or under-sampled regions of the family space.

Structural class is the second axis. Selection covers all-α, all-β, α/β, and $\alpha+\beta$ folds in approximately equal counts, with assignments drawn from CATH or SCOP architectural classification at the time of pre-registration. The reason is that structural class correlates with the inductive bias of conservation models; predominantly-helical proteins and predominantly-sheet proteins have systematically different conservation patterns at functional sites, and the framework's apparatus must be evaluated across this variation rather than within a structurally-homogeneous slice.

Functional class is the third axis. Selection covers enzymatic, structural, regulatory, and signaling proteins. The functional categorization is drawn from existing Gene Ontology annotations at the time of pre-registration. The reason is parallel to structural class; functional class correlates with the location and density of coherence-bearing positions (catalytic residues for enzymes, interface residues for structural proteins, allosteric sites for regulatory proteins, recognition residues for signaling proteins), and the framework's apparatus must be evaluated across this variation.

Catalytic-versus-non-catalytic position coverage is the fourth axis. Approximately half the selected families must be enzymatic with documented catalytic residues annotated in UniProt or PDB; the other half must include non-enzymatic functional sites (substrate-binding pockets, protein-protein

interfaces, regulatory sites). The reason is that catalytic residues are the cleanest case of documented coherence-bearing positions, and a family selection biased entirely toward enzymes would over-fit the M5 partition to one type of structural role.

Selection is committed at pre-registration with the exact list of Pfam family identifiers, the structural and functional class assignments per family, and the catalytic-residue annotations where applicable. Post-hoc family addition is excluded; post-hoc family removal is excluded except under the pre-registered exclusion criteria (failure of MSA quality control, ortholog-count drop below threshold after redundancy filter applied at locked parameters, family deprecation in the active Pfam release at analysis time)

6.2 Stream construction

Multiple sequence alignments are obtained from the active Pfam release, processed under standard MSA quality control, and converted to position-by-ortholog matrices over the 21-letter alphabet (20 amino acids plus gap). For each family, the matrix has dimensions $(L_{\text{family}} \times N_{\text{orthologs}})$, where L_{family} is the alignment length after low-coverage column trimming and $N_{\text{orthologs}}$ is the post-redundancy ortholog count.

Features in this domain are alignment positions. The symbol stream runs across orthologs at each position rather than across time within a single sequence. This is a structural adaptation of the induction pipeline, and the adaptation matters. The recursive-recoverability operationalization in D_1 measures temporal recoverability, whether structure at time t reappears as enabling structure at time $t + k$, while the operationalization in D_2 measures evolutionary recoverability, or whether structure at one ortholog reappears as enabling structure across the ortholog set. The framework's commitment to recursive structural reinforcement does not require temporal reinforcement specifically; it requires reinforcement across the iterations the substrate's dynamics produce. Evolutionary iteration is reinforcement in this sense, and adaptation is licensed by the framework's commitment rather than imposed against it.

The 5×3 grid runs identically on the position-by-ortholog matrices as it runs on the temporal streams in D_1, with the index that estimators traverse being ortholog rather than time. K_1 through K_4 accept the structural change without modification. A dictionary compresses the symbol sequence regardless of what the index means, n-gram MDL operates over conditional context regardless of whether context is temporal or evolutionary, the prequential transformer trains on the sequence without semantic commitment to the index's interpretation, and HMM/SSM MDL recovers latent

state structure across the index. K_5's factor enumeration likewise operates on the sequence without temporal interpretation. Operators A_1 through A_3 ablate positions rather than time-steps, which is the natural transposition of their D_1 operation.

6.3 Persistence operationalization

R1 in D_2 requires a per-position persistence measure. The natural measure is mutation tolerance, or how readily a position accepts substitution under the selection pressures the orthologs collectively expose. Position-wise mutation tolerance is computed as observed substitution frequency across the ortholog set after phylogenetic correction, where the correction accounts for non-independence due to shared ancestry through standard methods (independent-contrasts or maximum-likelihood ancestral-reconstruction approaches, with the specific method committed at pre-registration).

Observed substitution rate is itself a compression. It is the evolutionary-coherence compression; the substrate's persistence dynamics, integrated across the ortholog phylogeny, register as the rate at which each position accepts variation while the family remains functionally viable. The framework's commitment to validation as coherence-with-coherence comparison rather than coherence-against-truth is exactly satisfied by this construction. R1 in D_2 tests whether high induced w corresponds to low observed substitution tolerance, whether the framework's compression and evolution's compression agree on which positions persist as the family adapts.

The R1 statistic per family is the rank correlation (Spearman) between induced w and inverse mutation tolerance, with bootstrap confidence intervals constructed by ortholog resampling. The R1 pass at the global threshold from Section 8 requires the cross-family aggregate to clear the threshold under multiple-comparison adjustment.

6.4 Pfam as comparison compression, not adjudicator

Pfam's conservation annotations are the most extensively documented compression of the same evolutionary substrate. They are the accumulated product of biology's prevailing inductive biases; community-curated alignments, evolutionary conservation models drawn from substitution matrices and probabilistic profiles, structural homology arguments grounded in solved structures and homology modeling, and decades of expert curation that integrates experimental evidence with computational inference. The framework cannot accept Pfam as ground truth, and Section 4 already named the reason. Deference to incumbent symbolic authority is the failure mode the corpus

diagnoses, not a methodological convenience the corpus permits when the authority is well-credentialed.

Pfam enters D_2 as one more compression among several. Pfam's per-position conservation scores are processed through a parallel functional that takes Pfam's curation outputs as the relevance score and produces a Pfam-derived w-equivalent over positions. The Pfam-derived w-equivalent is then compared against the induced w produced by each (K, A) cell of the framework's grid. Convergence with Pfam contributes one additional R2 instance, the framework's compression and biology's prevailing compression agree on what the substrate exposes. The agreement is informative, meta-coherence with the most credentialed external compression of the same evolutionary data, but Pfam does not arbitrate. Cells where induced w diverges from Pfam are reported as informative findings, not as failures of induced w. Divergence may indicate that the framework's apparatus is detecting structure Pfam's curation conventions miss, or that Pfam is detecting structure the framework's apparatus misses, or that the two compressions are tracking partially distinct aspects of the same evolutionary substrate. Each reading is a finding; none positions Pfam as judge.

6.5 Intervention specifications

R3 in D_2 requires paired structural and interpretive interventions, with the structural-versus-interpretive distinction inherited from the framework's substrate-symbol commitment.

The structural intervention is in silico relaxed-selection simulation. The ortholog phylogeny is preserved, but substitution rates are uniformly increased through standard phylogenetic-perturbation methods (specific method committed at pre-registration, drawn from existing relaxed-clock or scaled-rate-matrix protocols); the resulting simulated ortholog set has the same phylogenetic structure but reduced selection pressure, which is a substrate change in the framework's sense, Ω-modifying because the constraint topology that selects against substitution at coherence-bearing positions has been relaxed. The induction pipeline is rerun on the perturbed ortholog set; $||\Delta w||$ between original and perturbed induced w is computed in L1.

The interpretive intervention is amino-acid relabeling under a permutation that preserves all BLOSUM-equivalent substitution patterns. The 20 amino acids are partitioned into BLOSUM-equivalence classes (groups whose pairwise BLOSUM scores fall within a fixed band), and a permutation is applied that maps each class to itself under arbitrary within-class shuffling. The result is a relabeled dataset that is statistically indistinguishable from the original under any analysis that respects BLOSUM relationships, a pure recoding that leaves biochemical relationships intact at the

equivalence-class level. The induction pipeline is rerun on the relabeled dataset; $\|\Delta w\|$ between original and relabeled induced w is computed.

The R3 ratio is structural $\|\Delta w\|$ divided by interpretive $\|\Delta w\|$, computed per family and aggregated across families. The R3 pass at the global threshold from Section 8 requires the aggregate ratio to clear the threshold under multiple-comparison adjustment.

6.6 D_2 feature-class partition for M5

The feature-class partition for M5 evaluation in D_2 is committed at pre-registration per family and is bound to documented biology rather than to induced w.

The coherence-bearing class consists of documented catalytic residues (where applicable), fold-defining positions (positions identified as load-bearing for the fold's structural integrity in the existing structural-biology literature), and conserved cores (operationalized as positions in the top conservation quintile across the ortholog set under standard conservation scoring, with the specific scoring method committed at pre-registration). The union of these three categories constitutes the coherence-bearing class for that family.

The noise class consists of surface-exposed non-functional residues (positions identified as solvent-accessible and not annotated to any functional role in existing structural and functional databases) and positions in the bottom conservation quintile under the same scoring. The union constitutes the noise class.

The partition is documented rather than derived. None of its components depends on induced w. Documented catalytic residues are drawn from UniProt and PDB annotations; fold-defining positions are drawn from existing structural-biology literature on each family; conservation quintiles are computed from standard scoring methods that predate the framework. The partition's epistemic status is constrained by documented biology, which is exactly what D_2's structural-transparency tier supplies.

M5 evaluation, per Section 2.5, computes mean induced w over the coherence-bearing class divided by mean induced w over the noise class, per (K, A) cell, per family. The per-family ratios are aggregated across families to produce the D_2 entry in the cross-domain consistency check at the M5 gate in Section 8. The aggregation respects per-family weighting (each family contributes equally) rather than per-position weighting (which would over-represent large families).

D_2 is now adapted to evolutionary dynamics, with its M5 partition fixed by documented biology and its R3 interventions calibrated to the structural-versus-interpretive distinction at the level of

selection pressure versus biochemical relabeling. The final and most exposed test is institutional, where the substrate is opaque, the symbolic layer is dense, and the framework's diagnostic of symbolic lock-in faces direct empirical challenge in the domain where the diagnostic originated. The next section specifies D_3.

Domain D_3: FOMC Statements with Multi-Level Coarse-Graining

D_3 is the opaque-substrate case. The generator, institutional decision-making at the Federal Open Market Committee, operating through deliberation, political constraint, and the deliberate ambiguity that characterizes central-bank communication, is inferable only through the stream itself and through whatever historiographic apparatus the analysis admits. The substrate is not specifiable in the way D_1's is; it is not even partially formalizable in the way D_2's is. What is observed is a corpus of public statements, and inferences about the policy substrate that produced them are themselves compressions running on the symbolic surface. The framework's claim is that induced w can distinguish substrate change from symbolic recoding even in this regime, where the two are most intertwined and most consequential. This section specifies the corpus, the three-level coarse-graining design, the persistence operationalization, the intervention specifications, the pandemic split, and the feature-class partition for M5.

7.1 Corpus construction

The primary corpus consists of FOMC statements from January 1994 through December 2019. The 1994 start is structural; it is the date the FOMC began issuing post-meeting statements as standard practice, making earlier periods incompatible with the corpus's structural assumption that each meeting produces a public statement. The December 2019 cutoff is the boundary of the pre-pandemic regime, defended in 7.5. The primary corpus contains approximately 200 statements, including both regularly-scheduled meeting statements and inter-meeting statements where issued. Statements from January 2020 onward are held as out-of-sample replication, providing approximately 50 additional statements through the analysis date. The replication set serves a separate evaluation function specified in 7.5.

The corpus is assembled from the Federal Reserve's official statement archive, with the canonical published text used in all cases. Pre-registered exclusion criteria are limited to statements issued under emergency circumstances where the standard statement structure was abandoned (no examples in the primary corpus, by inspection at pre-registration), statements that the Federal

Reserve subsequently amended where the original issuance is unrecoverable (no examples), and any statement whose canonical text is in dispute (no examples). Inclusion is otherwise exhaustive within the date window.

7.2 Three coarse-graining levels

The corpus is processed at three coarse-graining levels in parallel, with the full 5×3 grid run independently at each level. The three levels are designed to span the resolution range over which symbolic structure operates in institutional language, from sub-word tokens through syntactic chunks to semantic clusters.

Level 1 (token) uses byte-pair encoding tokenization at vocabulary size approximately 5,000. The tokenization is fitted on the primary corpus and held fixed for the replication set. Features at Level 1 are individual BPE tokens. The alphabet is large enough to capture meaningful sub-word structure (rate-decision tokens, regime-indicator tokens) without exploding into per-statement-unique vocabulary.

Level 2 (n-gram) uses 4-gram syntactic chunks over the Level 1 tokenization, with vocabulary size approximately 50,000 after frequency-based filtering of n-grams below a pre-registered occurrence threshold. Features at Level 2 are 4-gram chunks. The alphabet is calibrated to capture phrasal structure, the level at which forward-guidance language and balance-sheet idioms operate as compositional units rather than as individual tokens.

Level 3 (sentence-cluster) embeds each sentence in the corpus using a fixed pre-trained encoder (specific encoder committed at pre-registration), then clusters the embeddings via k-means at $k = 200$. Features at Level 3 are sentence-cluster identifiers. The alphabet size is small enough that each cluster captures a meaningful semantic role (rate-decision sentences, economic-assessment sentences, forward-guidance sentences, procedural sentences) without collapsing distinct semantic functions into a single cluster.

The full 5×3 grid is run at each of the three levels independently, producing 45 induced-w distributions per statement (or per analysis unit, depending on the persistence operationalization in 7.3). The cross-level comparison constitutes a within-domain coarse-graining stability test, exercising M3 from the prior admissibility commitments, the requirement that weight rankings be preserved under benign coarse-graining up to monotone transformation. Cross-level comparison tests whether high-w features at fine resolution map predictably to high-w features at coarser resolution.

M3 (within-domain coarse-graining stability) is distinct from M5 (across-domain translation invariance). D_3's coarse-graining sub-experiment exercises M3 specifically, and the M3 result is reported as a within-D_3 finding. M5 evaluation, in contrast, draws only on the Level 1 (token) results from D_3, where the 5×3 grid is run identically to D_1 and D_2. The Level 1 cut feeds the M5 cross-domain consistency check at the gate in Section 8; Levels 2 and 3 contribute to M3 within D_3 but do not enter M5.

7.3 Persistence operationalization

R1 in D_3 requires a post-statement persistence measure. The natural measure is regime stability following the statement's release. A statement that produces or coincides with subsequent regime breakdown, sharp moves in inflation expectations, yield-curve dislocations, volatility spikes, registers as low-persistence; a statement that produces or coincides with sustained calm registers as high-persistence.

Persistence per statement is operationalized as a composite of three financial-market measures evaluated over the 10 trading days following statement release. 1-month forward inflation-expectations volatility derived from TIPS breakeven rates, yield-curve slope volatility computed as the standard deviation of the 10-year minus 2-year Treasury spread, and VIX response measured as the average daily VIX deviation from its trailing 30-day mean. The three measures are standardized within the primary corpus (z-scored against the corpus distribution), then combined under fixed weights committed at pre-registration. The composite is signed such that higher composite values indicate lower persistence, larger volatility responses correspond to less stable post-statement regimes.

R1 tests whether high-w features in high-persistence statements are systematically distinct from high-w features in low-persistence statements. The R1 statistic is computed by partitioning the corpus into terciles by composite persistence score, computing the mean induced-w distribution within each tercile, and measuring the divergence between the high-persistence and low-persistence tercile distributions (under a pre-registered divergence measure, Jensen-Shannon at the feature level, with bootstrap CIs constructed by statement resampling). The R1 pass at the global threshold from Section 8 requires the divergence to clear the threshold under multiple-comparison adjustment across the three coarse-graining levels.

7.4 Intervention specifications

R3 in D_3 requires paired structural and interpretive interventions, with the structural-versus-interpretive distinction defended through documented FOMC history rather than through synthetic perturbation.

Structural interventions are documented monetary regime breakpoints in the primary corpus. Five breakpoints are pre-registered: the Greenspan-to-Bernanke chairmanship transition (February 2006), the QE1 announcement (March 2009), the introduction of explicit forward-guidance language tied to economic thresholds (August 2011), the rate-liftoff statement ending the zero-lower-bound regime (December 2015), and the introduction of the post-2012 inflation-targeting framework (January 2012). Each breakpoint is a documented substrate-level change in monetary policy regime, drawn from existing institutional-economics historiography on Fed policy.

Interpretive interventions are documented language-convention shifts that did not coincide with substrate-policy changes. Two are pre-registered as paradigm cases: the introduction of the "considerable period" formulation (August 2003), which represented a forward-guidance language convention shift without a coincident policy regime change, and the introduction of the "patient" formulation (January 2014), which similarly represented a language-convention shift within a stable policy regime. Additional language-convention shifts are identified through documented FOMC linguistic history (drawn from existing scholarship on Fed communication conventions), with the full pre-registered list published alongside the corpus and the structural-breakpoint list.

The R3 ratio is $||\Delta w||$ across structural breakpoints divided by $||\Delta w||$ across interpretive shifts, computed per (K, A) cell at Level 1 tokenization (the level on which M5 also operates) and aggregated across cells. Δw at a breakpoint is computed as the L1 difference between induced w averaged over a pre-registered window before the breakpoint and induced w averaged over a matched window after; the windows are size-matched to ensure the comparison is structural rather than artifact-of-window-length. The R3 pass threshold is 3:1 (structural to interpretive), with bootstrap CI requirements specified in Section 8.

7.5 The pandemic split

The decision to hold 2020+ as out-of-sample replication is methodologically necessary, not procedurally cautious. The pandemic regime introduces simultaneous structural and linguistic shifts that confound the structural-versus-interpretive distinction R3 depends on. The Federal Reserve's response to the pandemic combined a substrate-level policy regime change (zero-rate return,

balance-sheet expansion, new lending facilities) with a coincident language-convention shift (revised forward-guidance formulations, new framing for the Fed's economic assessment, restructured statement architecture). Including the pandemic period in the primary corpus would mean treating an event in which structural and interpretive interventions are confounded as data for a test that requires the two to be separable

The confounding is not analytically separable post-hoc, because the policy and language changes were simultaneous and mutually justified within FOMC communication. The policy change was framed in part through the language change, and the language change was authorized in part by the policy change. R3's pre-registered pass condition (structural $\|\Delta w\|$ at least three times interpretive $\|\Delta w\|$) requires a clean comparison between cases where only the substrate changed and cases where only the language changed. The pandemic period supplies neither and contaminates both.

Holding 2020+ as out-of-sample serves three functions. It preserves the cleanness of the primary R3 test by excluding the confounded period from the breakpoint analysis. It provides a separate post-hoc evaluation surface, where the framework's apparatus can be applied to the pandemic period as replication and the results examined for consistency with the primary findings under the explicit understanding that the pandemic data does not contribute to the R3 verdict. And it tests, in a structurally honest way, whether the framework's primary finding generalizes to a regime that is qualitatively different from the regimes it was tested on without conflating that generalization test with the original within-regime validation.

7.6 D_3 feature-class partition for M5

The feature-class partition for M5 evaluation in D_3 is committed at pre-registration through a curated FOMC-specific token-class lexicon, published alongside the pre-registration document and not adjustable post-hoc.

The coherence-bearing class consists of tokens documented in existing institutional-economics literature as substrate-policy-bearing: rate-decision tokens (the specific lexical units that mark the federal funds target language), balance-sheet tokens (the lexical units that mark asset-purchase and balance-sheet-policy language), regime-shift indicators (the lexical units associated with documented policy framework changes, including the inflation-targeting framework and the average-inflation-targeting framework where applicable), and dual-mandate tokens (the lexical units that explicitly invoke the Fed's statutory employment-and-price-stability mandate). The lexicon entries

are drawn from existing scholarly analyses of FOMC communication, with each entry citing the source literature.

The noise class consists of conventional discourse markers (the procedural and connective lexical units that appear in nearly every statement regardless of policy content), procedural language (committee-mechanics terminology that describes how the meeting was conducted rather than what was decided), calendar and date tokens, and committee-membership references (lexical units identifying meeting attendees and voting alignments at a procedural rather than substantive level).

The partition is documented rather than derived. Its components do not depend on induced w; they are drawn from existing institutional-economics scholarship and committed at pre-registration with the source lexicon published as part of the pre-registration package. M5 evaluation, per Section 2.5, computes mean induced w over the coherence-bearing class divided by mean induced w over the noise class, per (K, A) cell at Level 1 tokenization, aggregated across the primary corpus. The aggregate ratio is the D_3 entry in the cross-domain consistency check at the M5 gate in Section 8.

7.7 Why D_3 carries the most weight

D_3 is the domain where institutional substrate is most opaque, where the symbolic layer is densest, and where the framework's central diagnostic faces direct empirical examination. The corpus's claim about symbolic lock-in, that symbolic structure can persist even after decoupling from the substrate it was built to track, and that the decoupling is detectable through structural rather than purely interpretive analysis, is operationalized as R3 in D_3 at the level the claim originates. The framework either inherits the structural-versus-interpretive distinction at this domain's level of opacity, or it does not.

If R3 fails in D_3 while passing in D_1 and D_2, the framework's measurement operation works on transparent and structurally-transparent substrates but loses its substrate-symbol distinction at the institutional-opacity tier. The claim about institutional dynamics is exposed at the level where it is most consequential, and the corpus's universality claim is correspondingly restricted. If R3 holds in D_3, the framework has empirical purchase on a question, whether central-bank communication exhibits structural change distinguishable from linguistic-convention shift, that classical text-analysis methods cannot answer because they lack the substrate-symbol architecture the framework supplies. The result, in either direction, registers a finding that the corpus binds to specific consequences in the next section.

With all three domains now specified, including their M5-relevant feature-class partitions and their R1, R2, R3 operationalizations, the empirical program is fully grounded at the domain level. The next section binds the program to interpretive commitments, pre-registering what each possible result configuration means for the framework before any data is touched, and binding M5's gate logic to the R-signature outcome matrix.

The Outcome Interpretation Matrix, the M5 Gate, and Pre-Registration Discipline

The empirical program is now fully specified at the domain level. What remains is to bind every possible result configuration to a specific framework consequence in advance. Without this binding, the program is a measurement specification; with it, the program is a commitment device. This section sequences the M5 admissibility gate before the eight-cell outcome matrix, specifies what each cell means under both M5-pass and M5-fail, flags the structurally-interesting failure configurations, commits to joint distribution reporting, and locks the statistical thresholds in formal block.

8.1 Why pre-registration is non-optional here

The framework already faces the challenge of unfalsifiable universality. The challenge is not unreasonable on its face; a coherence-first ontology that defines the real through persistence-under-arbitration and rejects correspondence-truth as a category error invites the suspicion that whatever the empirical results turn out to be, the framework will absorb them. Post-hoc interpretation of partial results would confirm that suspicion in operational form. The framework would have produced an empirical program, run it, and then read each result as compatible with its claims, which is exactly the symbolic-lock-in pattern the framework names as failure.

Pre-registration is not methodological hygiene imported from open-science conventions. It is structural integrity for a framework whose central commitment is to test itself in its own terms. The framework's claim is that it can be wrong in ways the framework can name in advance. The matrix that follows is the operational form of that claim. Each cell is a configuration the framework could exhibit, and each cell is bound to a specific consequence the framework accepts before any data is run. The binding is non-negotiable post-hoc. The framework either holds itself to it or fails the same coherence audit it imposes on every other symbolic system in the corpus.

8.2 The M5 admissibility gate

M5 is evaluated first. The gate sequencing is structural, not procedural. M5 tests whether the induced-*w* functional is the same functional across domains, and that test is logically prior to the within-domain measurement tests, because the within-domain tests presuppose that what is being measured is one thing rather than three different things sharing a name.

M5-pass requires the cross-domain ratio of mean induced *w* over coherence-bearing features to mean induced *w* over noise features, after rank-normalization within each domain, to be within a factor of 2 across all three domain pairs (D_1 /D_2, D_1 /D_3, D_2/D_3) for at least three of the five estimator classes. The 3-of-5 requirement matches the cross-philosophy convergence discipline at the within-domain layer. The same logic that authorizes R2, agreement across philosophically-decoupled estimator philosophies, authorizes M5, requiring philosophical agreement on the cross-domain consistency of the functional itself. The factor-of-2 threshold under rank-normalization is calibrated to admit functionals consistent up to monotone transformation while rejecting functionals that drift across domains in ways the monotone-transformation clause cannot absorb.

M5-pass licenses the eight-cell matrix interpretation at full strength. The framework's universality claim survives the admissibility check, and the matrix's reading of within-domain results carries the universality claim with it.

M5-fail does not void the matrix. The matrix is computed and reported in either case. M5-fail conditions the matrix's reading. Whatever the framework does within domains, it does as a different functional in each domain, and the universality claim is foreclosed at the functional level regardless of how the within-domain signatures resolve. The M5-fail interpretation is specified separately in 8.4, and it is informative. Different patterns of M5-fail correspond to distinct revision pathways for the corpus.

The gate is not optional. It is checked first, and its result conditions everything below.

8.3 The eight-cell outcome matrix under M5-pass

Three binary signatures yield 2^3 = 8 outcome configurations. Each is bound in advance to a specific framework consequence under M5-pass.

R1✓ R2✓ R3✓ — full pass. The framework is internally consistent under the tested regime. Induced *w* predicts persistence, converges across philosophically-decoupled estimators, and discriminates structural intervention from interpretive intervention. Universality is conditional on the tested K, A, and domain classes; the substrate is validated as stipulated. The framework has earned operational

status within the regime tested. The corpus's universality claim is empirically grounded at the level of the program's coverage.

R1✓ R2✓ R3✗. Induced w predicts persistence and converges across philosophies but does not distinguish structural from interpretive intervention. The framework's diagnostic apparatus, the substrate-symbol distinction, is exposed at its operational layer, and the symbolic-lock-in claim is weakened. The framework can detect what persists and can do so reproducibly, but cannot tell substrate change from pure recoding, which is exactly what the lock-in diagnostic requires. The corpus's diagnostic of institutional dynamics is restricted; what remains is a coherence-tracking apparatus without the structural-versus-symbolic resolution the corpus claims it has.

R1✓ R2✗ R3✓. Induced w predicts persistence and discriminates intervention classes but is estimator-dependent. Convergence fails across philosophical decoupling. Universality is false at the convergence layer; conditional validity within a chosen estimator philosophy is true. The framework operates as a coherent measurement apparatus within whichever estimator philosophy the analyst commits to, but the choice of philosophy matters in a way the universality claim does not permit. The corpus's universality claim is replaced by a conditional-validity claim, which is a substantive restriction.

R1✓ R2✗ R3✗. Induced w predicts persistence but lacks both convergence and intervention asymmetry. The framework's "structural" notion collapses to "predictive of persistence under this estimator." This is the configuration in which the framework reduces to standard predictive-modeling. It produces a useful predictive apparatus but cannot claim to be tracking structure in the framework's specific sense. The corpus's distinguishing claims, universality, substrate-symbol distinction, both fail. What remains is consistent with classical predictive-modeling traditions; the framework's contribution above those traditions is voided.

R1✗ R2✓ R3✓. Induced w converges and discriminates intervention classes but does not predict persistence. The framework names something stable across philosophical decoupling, and that something inherits the substrate-symbol distinction, but it does not connect to the framework's own ontological commitment. The configuration is the most diagnostically interesting failure mode the program admits. The framework would have correctly diagnosed an ontologically empty closure, a stable, internally-consistent symbolic structure decoupled from the persistence dynamics that authorize it, and then exemplified it. The corpus must revise; either the persistence operationalization is wrong (and the empirical layer needs reconstruction), or the ontological commitment to persistence-as-criterion is wrong (and the philosophical layer needs

reconstruction), or the inference from "induced *w* names what coheres" to "induced *w* names what persists" was always weaker than the corpus claimed.

R1✗ R2✓ R3✗. Induced *w* converges only. The measurement operation is self-referentially closed across philosophically-decoupled estimators, but neither predicts persistence nor distinguishes intervention class. The framework has produced symbolic lock-in of itself, a stable measurement consensus that does not track what the framework says it tracks. The configuration is the cleanest case of the corpus's central diagnostic applied to the corpus, and it would require fundamental revision at the level where the inference pipeline produces its outputs.

R1✗ R2✗ R3✓. Induced *w* distinguishes intervention classes but neither predicts persistence nor converges across philosophies. The framework reduces to a perturbation-class detector, a tool that responds differently to structural and interpretive perturbations under whichever estimator was chosen, without the universality and ontology that the framework claims accompany the response. The corpus's contribution above existing intervention-asymmetry methods becomes uncertain.

R1✗ R2✗ R3✗ — empirical void. None of the signatures pass. Universality, predictive content, and the substrate-symbol distinction all fail. The framework requires fundamental revision at every layer the program tests. The corpus does not survive the result intact.

8.4 The eight-cell outcome matrix under M5-fail

If M5 fails, the universality claim is foreclosed at the functional level, but the within-domain results remain informative. The matrix is computed identically; its reading shifts.

M5-fail with R-signature pass in D_1 only. The framework operates as a structured-stream measurement apparatus on synthetic substrates whose generators are written. Within D_1, the measurement operation is internally consistent across philosophically-decoupled estimators, predicts persistence, and discriminates intervention classes. Outside D_1, the apparatus fails one or more of those tests, and M5 reports that the failures are not just measurement failures but functional differences. The corpus's universality across natural substrates fails. What remains is a tool for analyzing controlled streams, which is a contribution but not the contribution the framework claims.

M5-fail with R-signature pass in D_1 and D_2 but not D_3. The framework operates on substrates with at least structural transparency. The opaque-substrate tier, where the framework's institutional-dynamics claim lives, is where the apparatus breaks down. The corpus's claim about institutional symbolic lock-in is restricted; the diagnostic that originated in the analysis of opaque institutional symbolic systems cannot, on this result, be operationalized in those systems through induced *w*. The

framework retains coherence on transparent and partially-transparent substrates and loses it where its most consequential claims live.

M5-fail with R-signature pass across all three domains. The framework's measurement operation works within each domain individually, but the functional varies across domains. This is the scope-pluralism configuration. The framework names a class of measurement operations rather than a single one, and the within-domain coherence of each member of the class is real, but their cross-domain identity is not. The universality claim is replaced by a pluralism claim, and the corpus must revise the universality language throughout. What remains is a family of coherent measurement operations, each appropriate to its substrate-transparency tier, with shared structural logic but distinct functional implementations. This is a substantive contribution that the framework had not previously named as its actual scope.

Each M5-fail configuration is a distinct revision pathway. M5-fail is not catastrophic. It is informative at a different layer than R-signature failure, and the corpus accepts the layer-specific reading in advance.

8.5 Structurally-interesting failures

Three configurations are flagged as structurally-interesting failures; cells where the framework would correctly diagnose a failure mode it has previously named in the corpus and then exemplify it. These are not "bad" results in the conventional sense. They are the framework auditing itself in its own terms, which is the recursive coherence the corpus claims throughout.

R1✗ R2✓ R3✗ exemplifies symbolic lock-in of the framework's measurement operation; stable across philosophies, disconnected from persistence and from the substrate-symbol distinction. The corpus's diagnostic applied to itself, registered as the result the framework most needed to be able to detect.

R1✗ R2✓ R3✓ exemplifies ontologically-empty closure; stable, structurally-discriminating, but disconnected from the ontological commitment that authorizes the structure-naming. The corpus's distinction between "internally consistent" and "tracking what the ontology says it tracks" registered as an empirical separation.

M5-fail with universal R-signature pass exemplifies scope-pluralism; each domain's measurement operation works, but the operation is not one operation. The corpus's universality claim registered as a claim the framework had been making about itself without operational basis.

In each case, the framework's pre-registered architecture would name the failure mode the result instantiates, and the corpus accepts that the result, by registering as that failure mode, supplies a contribution the framework can absorb. The matrix is honest about which failures hurt the framework most and which constitute the framework's most informative self-corrections.

8.6 Joint and marginal reporting

The four conditions are not statistically independent. Perturbations underlie both R1 and R3, though the perturbations differ structurally, and bootstrap covariance between the two is non-trivial. Convergence (R2) depends on data structure that may correlate with persistence (R1) through shared substrate features. M5 ratios depend on the same induced-w distributions that all three R-signatures interrogate. Reporting only marginal pass/fail would discard structural information about the framework's behavior.

The analysis reports joint distributions across the {K × A × Domain} grid for all four conditions. If the correlations among signatures are themselves stable across the grid, if M5 and R2 covary in the same way under K_1 as under K_5, if R1 and R3 covary in the same way in D_1 as in D_3, that stability is a meta-coherence finding worth flagging in its own right. The framework's inference pipeline produces correlated outputs in structurally consistent ways, which is its own form of internal consistency. If the correlations vary erratically across the grid, that is also a finding. The framework's outputs covary in ways that depend on which estimator-operator-domain combination produced them, which exposes a layer of contingency the framework's universality claim does not permit.

8.7 Statistical thresholds, locked

The following thresholds are committed at pre-registration and are not adjustable post-hoc.

M5 pass: cross-domain mean-w ratio (coherence-bearing / noise), rank-normalized within domain, within a factor of 2 across all three domain pairs (D_1 /D_2, D_1 /D_3, D_2/D_3) for at least 3 of 5 estimator classes; bootstrap lower 95% CI on the cross-domain consistency measure must clear the pre-specified consistency threshold.

R2 pass: cross-philosophy median Spearman rank correlation > 0.6, with lower 95% CI > 0.4.

R1 pass: Cohen's d > 0.5, with lower 95% CI > 0.3, at $p < 0.01$ after multiple-comparison adjustment.

R3 pass: structural-to-interpretive $||\Delta w||$ ratio > 3.0, with lower 95% CI > 2.0.

Multiple-comparison adjustment: Bonferroni across all 51 cells of the primary {K × A × Domain} grid (5 estimators × 3 operators × 3 domains, with the D_3 Level 1 tokenization as the M5-relevant cut) plus the D_3 coarse-graining sub-experiment cells, plus the M5 cross-domain comparisons.

Bootstrap procedure: B = 1,000 replicates throughout. Time-window resampling at block length equal to one mean regime sojourn for R2 and M5 in temporal domains; ortholog resampling for R2 and M5 in D_2; statement resampling for R2 and M5 in D_3. Standard non-parametric bootstrap for R1 and R3 statistics.

The thresholds are calibrated to admit substantively meaningful effects while rejecting effects that could plausibly arise from noise. They are stricter than conventional psychology-replication thresholds and looser than physics-discovery thresholds, which is appropriate for a program that occupies an intermediate epistemic register, an empirical specification of a theoretical framework's measurement operation, evaluated under controlled and partially-controlled conditions.

With outcomes pre-bound to interpretation and the M5 gate sequenced before the matrix, the program is now structurally complete. The remaining work is to position the paper within the corpus, articulate what the framework would gain or lose under each result, and hand the program off to whoever runs it.

Implications, Handoff, and What This Paper Does Not Do

The program is now complete as a specification. What remains is to articulate what the corpus gains under each result configuration, to register the architectural separation of admissibility from measurement as a contribution independent of any data, to acknowledge the handoff this paper performs, to state the limitations the framework's own commitments require, to address the critiques the paper most plausibly invites, and to close on the next question the substrate creates.

9.1 What the corpus gains under each result class

The matrix in Section 8 binds each configuration to a specific consequence. The aggregate effect is that the corpus emerges from the program with sharper boundaries regardless of how the result resolves.

A broad M5 pass empirically grounds the cross-domain admissibility commitment inherited from prior induced-weights work. That commitment, that induced *w* must be compatible across domain mappings that preserve coherence up to monotone transformation, is the deepest claim in the universality program, and it has until now been an admissibility constraint stated as a structural

requirement without empirical instantiation. A broad pass would establish that the constraint is satisfiable in practice, across the substrate-transparency gradient the program tests. That is the strongest cross-domain claim available under the framework's own ontology.

A broad R2 pass empirically grounds the convergence claim within the tested estimator and operator classes. Convergence across philosophically-decoupled compressors is what the corpus has needed to demonstrate that induced *w* tracks something exceeding the inductive biases of the family of estimators that compute it. A broad pass would establish that the substrate is responding to structural features that disagree-prone philosophies are nonetheless driven into agreement on, which is the operational signature the universality claim requires.

An R1 pass empirically grounds Adaptive Realism's persistence-as-criterion. The framework's ontological commitment, that the real is what persists across recursive coherence arbitration, has been a structural claim about the framework's terms. R1 would convert it into a measurable property of the framework's apparatus; high-*w* features participate in persistence in ways low-*w* features do not. The conversion does not establish realism in any external sense; it establishes that the framework's ontology and the framework's measurement operation cohere with one another, which is what the framework requires.

An R3 pass empirically grounds the symbolic-lock-in diagnostic. The corpus's central diagnostic distinction, between substrate change and pure recoding, between Ω-modifying intervention and dynamics-preserving relabeling, has been a structural distinction the framework has named without operationalizing in cross-domain empirical form. R3 would establish that induced w inherits the distinction across domains of differing transparency, and that the inheritance is detectable through the framework's apparatus rather than through correspondence-truth methods the framework rejects.

Each failure tightens the universality claim by characterizing exactly where the framework actually applies. The corpus does not lose under failure; it gains a sharper map of its own scope. That is the structure of the contribution under either direction of the result.

9.2 The M5/R-signature separation as a contribution in itself

Before any data is run, the architectural move performed in this paper is itself a contribution to the framework. The move is the separation of admissibility from measurement, and the recognition that the framework's universality claim is not a single empirical claim but two distinct claims operating at different layers.

The first claim is that the induced-*w* functional is the same functional across domains. That claim lives at the admissibility layer and is tested by M5. The second claim is that the functional behaves coherently within each domain. That claim lives at the measurement layer and is tested by R1, R2, and R3. The two claims can succeed and fail independently, and each failure mode admits a distinct revision pathway for the corpus.

The separation clarifies what universality means in the framework's own terms. Without the separation, a partial result would be ambiguous; a mixed pattern of within-domain successes and failures could be read as a measurement problem, a functional problem, or some combination, with no principled way to attribute the result to a layer. With the separation, a partial result is informative. M5's verdict and the within-domain matrix's pattern together specify which layer is exposed and what kind of revision the exposure requires. Future extensions of the corpus can target either layer independently, and failure modes can be diagnosed at the layer where they originate rather than at the level of the program as a whole.

The architectural separation is portable. Other empirical programs in the corpus, programs the next paper, or the paper after that, may specify, inherit the admissibility-versus-measurement distinction as a structural template. The contribution is not bounded by the data this program will produce.

9.3 Why this paper does not run the experiments

The paper performs theoretical work that compresses to operational closure. The induced-weights program now has a fully specified, falsifiable, cross-domain validation architecture, with admissibility and measurement layers cleanly separated, with feature-class partitions documented per domain, with statistical thresholds locked, with intervention specifications committed, and with the outcome interpretation matrix bound in advance. An engineer with access to the specified estimators, the specified operators, the specified domains, and the specified statistical apparatus can implement the program from this paper alone.

Running the program requires engineering capacity and compute resources I cannot assume in this paper. The handoff is the work. The pattern is the one EPR established for thought-experiment specification, the one Bell established for inequality specification, and the one any number of theoretical-specification papers have established since, a theoretical contribution that binds subsequent empirical work without requiring its author to perform that work.

This is not a limitation of the paper. It is the paper's mode. The corpus operates by recursive layering, where each paper compresses prior work to the level just above where the next paper begins, and

where the next paper may be written by the same author or by someone else who picks up the spec and runs it. The handoff is part of the architecture.

9.4 The recursive structure of the contribution

The corpus is layered self-extension. Each paper extends the substrate by compressing prior work to the level just above where the next paper begins.

Formal Foundations of Coherence Information Theory established coherence-weighted entropy and mutual information, along with capacity and selective compression theorems, conditional on a bounded symbol-weight function. *Engineering Induced Coherence Weights for Coherence Information Theory* specified the engineering interface, coherence proxies, ablation-induced relevance, weight derivation, coding under rate constraint, benchmark suite, and robustness harness. *Formal Foundation of Induced Coherence Weights* specified the derivation pipeline from observed streams to induced weights, the admissibility commitments M1 through M6, and the residual hard seam at the choice of estimator. This paper specifies the cross-domain validation architecture, separates admissibility (M5) from measurement (R1, R2, R3), and binds outcomes to interpretations in advance.

The next paper, written by whoever runs this program, should take results back into theory. It will tighten or revise universality claims at whichever layer the data exposes. If M5 passes and R-signatures pass, that paper sharpens the framework's universality language and identifies the boundary conditions the program's coverage did not test. If M5 fails or R-signatures fail, that paper revises the framework's commitments at the layer the failure exposes, and the corpus's structure absorbs the revision through the same recursive logic that produced this paper. The corpus is not a closed system. It is a substrate that extends itself through its own commitments, and this paper continues that mode.

9.5 Limitations, framework-compliant

This paper does not list defensive limitations of the conventional kind, because the conventional limitations would import the correspondence-truth framing the framework rejects. The acknowledgments below are structural; they name what the program does and does not establish in the framework's own terms.

Validation operates through coherence-with-coherence comparisons. The program does not test induced w against an external ground truth and does not claim to. This is consistent with Synthetic

Epistemology's foreclosure of correspondence-truth as a category error, and the program's design respects the foreclosure throughout.

Convergence among compressions, registered through R2, does not establish realism. It establishes that the measurement operation produces structurally stable outputs across philosophically-decoupled estimators within a domain. The framework's ontology authorizes the inference from convergence to "structure exceeds estimator," but does not authorize the inference from convergence to "structure exists independently of any compression."

M5 admissibility, in the program's operationalization, does not establish the existence of a domain-invariant ontological substrate. It establishes that the induced-*w* functional, as a measurement apparatus, recognizes the same kind of structural distinction across substrates of differing transparency. The functional's cross-domain consistency is a property of the apparatus and of its interaction with the substrates it analyzes, not a property of an underlying domain-invariant reality the framework would have no grounds to posit.

Persistence metrics, registered through R1, are themselves compressions of substrate dynamics. Mutation tolerance is the evolutionary-coherence compression in D_2; post-statement market volatility is the institutional-stability compression in D_3; regime-inference recovery is the dynamical-stability compression in D_1. Their use as persistence operationalizations is framework-compliant via Adaptive Realism's commitment that persistence is operationally measurable, not via any claim that the chosen compressions exhaust what persistence is.

The transparency hierarchy across D_1 through D_3 is methodological scaffolding, not an empirical-versus-synthetic distinction. All three domains are formally synthetic in the framework's sense; all three are streams produced by generators whose mediation is acknowledged. The hierarchy is a structural axis along which the universality claim can be tested, not a quality ranking that privileges one domain over another.

9.6 Anticipated critiques

Four critiques are reasonable and are addressed directly.

The unfalsifiability critique. This framework, the critique runs, is unfalsifiable because its coherence-first ontology absorbs whatever empirical pattern the data produces. The critique is refuted by the M5 gate, the eight-cell outcome matrix, and the pre-registered statistical thresholds. The matrix names eight configurations under M5-pass and three additional configurations under M5-fail, with each bound to a specific framework consequence. Several of those configurations require the

framework to revise its central claims of universality, convergence, persistence, and substrate-symbol distinction at named layers. The framework can be wrong, and the program names exactly how, in advance.

The "defining your way out of validation" critique. The program, the critique states, replaces validation with a definitional structure that calls itself validation. The critique is refuted by three independent failure modes the program admits; cross-philosophy convergence (R2) can fail if estimator philosophies disagree on what the substrate exposes, intervention asymmetry (R3) can fail if induced w cannot distinguish substrate change from pure recoding, and cross-domain admissibility (M5) can fail if the functional varies across domains. Each failure mode has a well-defined statistical signature, each is evaluated under pre-registered thresholds, and each maps onto distinct revision pathways for the corpus. Definition without empirical content does not produce these failure modes; the program's structure is incompatible with the critique's premise.

The synthetic-data critique. Synthetic data, the critique argues, cannot validate claims about real systems. The critique is refuted by Section 4's transparency-hierarchy argument; all three domains are formally synthetic in the framework's sense, the synthetic-versus-empirical illusion is a residue of correspondence-truth thinking, and the program does not ask synthetic data to do anything the other domains are not also asked to do. D_1's transparency is its diagnostic asset, not a weakness; D_2 and D_3 supply substrate variation D_1 cannot, and the program tests the framework against the full gradient.

The M5-calibration critique. The M5 threshold, the critique runs, is either too lax (admitting functionals the universality claim should reject) or too strict (rejecting functionals the universality claim should admit). The critique is refuted by the explicit threshold specification and by the structurally-interesting M5-fail configurations in Section 8.4. The threshold is not a verdict knob; it is a calibration that produces informative outcomes at both pass and fail. M5-pass licenses universality; M5-fail with within-domain success registers as scope-pluralism, which is itself a substantive contribution. The threshold's calibration can be debated, but the program's structure does not depend on the threshold being uniquely defensible. It depends on the threshold being committed in advance and on its outcomes being bound to specific consequences.

9.7 Closing

The program either runs or it does not. If it runs, the framework either has empirical purchase, or it does not. In either case, the next theoretical question becomes specific in a way it has not been until

now. If M5 passes broadly and the R-signatures pass broadly, the next question is what the boundary conditions of universality are, what substrates lie outside the tested gradient, what estimator philosophies lie outside the tested class, and what failure modes the program's coverage did not expose. If M5 passes but R-signatures partially fail, the next question is what the framework actually measures, given that its functional is consistent but its measurement operation falls short within domains. If M5 fails with R-signatures passing per-domain, the next question is what the scope of the pluralism the framework has named actually is, what unifies the family of measurement operations, and what each member tracks within its domain. If signatures fail broadly, the next question is which substrate commitments need revising, and at what level, ontological, epistemic, operational, or all three.

The framework is constructed to make these questions answerable rather than to forestall them. The program specified here is the discipline that produces the answers, and the answers will produce the next paper, which will continue the corpus's recursive layering by compressing this paper's contribution to the level just above where the next layer begins. That is what the substrate does. That is what this paper does.

References

James, B. (2025). Beyond Shannon: Coherence Information Theory and the Future of Communication. PhilPapers. https://philpapers.org/rec/JAMBSC

James, B. (2026). Engineering Induced Coherence Weights for Coherence Information Theory. PhilPapers. https://philpapers.org/rec/JAMEIC-2

James, B. (2025). Formal Foundations of Coherence Information Theory: Capacity and Compression Theorems. PhilPapers. https://philpapers.org/rec/JAMFFO-2

James, B. (2026). Formal Foundation of Induced Coherence Weights: Compression-based Coherence and Operational Information Measures. PhilPapers. https://philpapers.org/rec/JAMFFO-3

James, B. (2025). Recursive Coherence: Foundations for an Adaptive Philosophy of Science, Mind, and Society. PhilPapers. https://philpapers.org/rec/JAMRCF

James, B. (2025). Synthetic Epistemology: Engineering Coherence Beyond Truth. PhilPapers. https://philpapers.org/rec/JAMSEE-2

Appendix A — Pseudocode for the Induced-Weights Pipeline

This appendix specifies the induced-weights pipeline at the level of executable pseudocode, with the 5×3 grid structure, the three intervention protocols, and the M5 cross-domain ratio computation made explicit. The pipeline is adapted from the engineering interface specified in *Engineering Induced Coherence Weights for Coherence Information Theory*, with modifications required by the cross-domain validation architecture committed in this paper. Notation follows the body of the paper. The pseudocode is language-agnostic; an engineer implementing the program in Python, R, Julia, or any modern scientific computing environment can map the structures directly.

A.1 Top-level orchestration

```
PROGRAM meta_coherence_validation:
    INPUT: domain_specs = {D1, D2, D3}        # per Sections 5, 6, 7
           estimator_set = {K1, K2, K3, K4, K5}  # per Section 3.1
           operator_set  = {A1, A2, A3}          # per Section 3.2
           thresholds = pre_registered_thresholds  # per Section 8.7
           seed = pre_registered_random_seed
    OUTPUT: meta_coherence_report

    initialize random state with seed
    initialize result_store empty

    # --- Phase 1: per-domain induction across the 5x3 grid ---
    FOR each domain d IN domain_specs:
        stream_d = load_or_generate_stream(d)
        partition_d = load_feature_class_partition(d)   # coherence-bearing
vs noise
        FOR each estimator K IN estimator_set:
            FOR each operator A IN operator_set:
                w_dKA = induce_weights(stream_d, K, A)
                store w_dKA in result_store[d][K][A]

    # --- Phase 2: within-domain R-signature evaluation ---
    FOR each domain d IN domain_specs:
        R2_d = evaluate_R2(result_store[d], thresholds.R2)
        R1_d = evaluate_R1(result_store[d], stream_d, d, thresholds.R1)
        R3_d = evaluate_R3(result_store[d], stream_d, d, thresholds.R3)
        store {R1_d, R2_d, R3_d} in signature_results[d]

    # --- Phase 3: M5 admissibility gate evaluation ---
    M5_result = evaluate_M5(result_store, partition, thresholds.M5)

    # --- Phase 4: outcome interpretation matrix binding ---
    interpretation = bind_outcomes_to_matrix(
        M5_result,
        signature_results,
        outcome_matrix_specification  # per Section 8.3 and 8.4
    )
```

```
    # --- Phase 5: joint distribution reporting ---
    joint_report = report_joint_distributions(
        result_store,
        signature_results,
        M5_result
    )

    RETURN meta_coherence_report = {
        signature_results,
        M5_result,
        interpretation,
        joint_report
    }
```

The orchestration computes the entire {K × A × Domain} grid before evaluating any of the four conditions, which is required by the joint distribution reporting commitment in Section 8.6. No condition is computed in isolation; all four read against the same underlying induced-w distributions.

A.2 The induce_weights subroutine

The subroutine instantiates the derivation pipeline from prior induced-weights work, with the proxy-form choice fixed at predictive-logloss per Section 3.4 and the secondary compression-delta sanity check applied only in D_1.

```
FUNCTION induce_weights(stream, K, A):
    INPUT: stream — the symbol/feature stream over the domain's index
           K      — estimator from {K1, ..., K5}
           A      — ablation operator from {A1, A2, A3}
    OUTPUT: w     — induced-weight distribution over features

    # Step 1: compute baseline coherence proxy under estimator K
    C_baseline = compute_coherence_proxy(stream, K,
form="predictive_logloss")

    # Step 2: compute relevance score per feature under operator A
    features = enumerate_features(stream)
    relevance = empty dictionary

    SWITCH A:
        CASE A1 (leave-one-out):
            FOR each feature x IN features:
                stream_ablated = ablate_feature(stream, x, mode="LOO")
                C_ablated = compute_coherence_proxy(stream_ablated, K,

form="predictive_logloss")
                relevance[x] = C_baseline - C_ablated

        CASE A2 (Shapley, k=64 sampled coalitions):
```

```
            FOR each feature x IN features:
                shapley_estimates = empty list
                FOR i FROM 1 TO 64:
                    coalition = sample_coalition_excluding(features, x)
                    stream_with    = ablate_feature_set(stream,
                                                        features \ (coalition
∪ {x}))
                    stream_without = ablate_feature_set(stream,
                                                        features \ coalition)
                    C_with    = compute_coherence_proxy(stream_with, K,

form="predictive_logloss")
                    C_without = compute_coherence_proxy(stream_without, K,

form="predictive_logloss")
                    append (C_with - C_without) to shapley_estimates
                relevance[x] = mean(shapley_estimates)

        CASE A3 (correlation-cluster group ablation):
            clusters = compute_correlation_clusters(stream, threshold=0.5,
                                                    metric="pearson")
            cluster_relevance = empty dictionary
            FOR each cluster c IN clusters:
                stream_ablated = ablate_feature_set(stream, c)
                C_ablated = compute_coherence_proxy(stream_ablated, K,

form="predictive_logloss")
                cluster_relevance[c] = C_baseline - C_ablated
            FOR each feature x IN features:
                c = cluster_containing(x)
                relevance[x] = cluster_relevance[c] / size(c)

    # Step 3: induce bounded weights via monotone bounded mapping
    w = empty dictionary
    FOR each feature x IN features:
        w[x] = sigmoid(beta * relevance[x])    # beta pre-registered

    # Optional D1-only sanity check on proxy form
    IF stream.domain == D1:
        relevance_alt = recompute_relevance(stream, K, A,
form="compression_delta")
        log_proxy_form_consistency(relevance, relevance_alt)

    RETURN w
```

The compute_coherence_proxy routine wraps the five estimator-specific implementations. K_1 calls a zstd interface and returns the negative compressed length normalized by stream length; K_2 runs arithmetic-coded n-gram MDL across $n \in \{3, 4, 5\}$ and averages; K_3 trains a small transformer with prequential coding and returns the cumulative negative log-likelihood; K_4 performs MDL search over hidden-state cardinality and emission structure and returns model-cost-plus-likelihood; K_5

implements the Kaspar-Schreiber factor enumeration and returns factor count. Implementation details for each are out of scope for this appendix; the relevant prior-work specification covers them.

A.3 R2 evaluation — within-domain cross-philosophy convergence

```
FUNCTION evaluate_R2(domain_results, threshold_spec):
    INPUT: domain_results — w distributions across all (K, A) cells in one
domain
           threshold_spec — R2 pass threshold (median Spearman > 0.6, lower
CI > 0.4)
    OUTPUT: R2_verdict — pass/fail with statistics

    cell_pairs = enumerate_pairs(domain_results)
    spearman_scores = empty list

    FOR each (cell_i, cell_j) IN cell_pairs:
        w_i = domain_results[cell_i]
        w_j = domain_results[cell_j]
        rho = spearman_rank_correlation(w_i, w_j)
        philosophical_class_i = classify_philosophy(cell_i)   #
coding/predictive/algorithmic
        philosophical_class_j = classify_philosophy(cell_j)
        cross_philosophy = (philosophical_class_i != philosophical_class_j)
        append (rho, cross_philosophy) to spearman_scores

    cross_philo_rhos = filter spearman_scores WHERE cross_philosophy
    median_rho = median(cross_philo_rhos)

    bootstrap_CIs = bootstrap_median_spearman(
        domain_results,
        B = 1000,
        method = "time_window_resample"     # for D1, D3
                 OR "ortholog_resample"     # for D2
    )

    R2_verdict = {
        median = median_rho,
        CI_lower = bootstrap_CIs.lower_95,
        passes = (median_rho > threshold_spec.median_threshold AND
                  bootstrap_CIs.lower_95 >
threshold_spec.CI_lower_threshold),
        within_philosophy_block_correlations = compute_block_correlations(
            spearman_scores, philosophical_classes
        )
    }
    RETURN R2_verdict
```

Within-philosophy block correlations are reported separately as the discriminative diagnostic Section 3.3 requires. High within-philosophy correlation paired with low cross-philosophy correlation would register as the family-internal-consistency failure mode named in Section 2.

A.4 R1 evaluation — persistence prediction

The operationalization differs by domain per Sections 5.5, 6.3, and 7.3. The pseudocode dispatches on domain.

```
FUNCTION evaluate_R1(domain_results, stream, domain, threshold_spec):
    INPUT: domain_results — w distributions across all (K, A) cells
           stream         — the domain's stream
           domain         — D1, D2, or D3
           threshold_spec — Cohen's d > 0.5, CI lower > 0.3, p < 0.01
    OUTPUT: R1_verdict — pass/fail with statistics

    SWITCH domain:
        CASE D1:
            persistence_per_feature = empty dictionary
            FOR each feature class fc IN stream.feature_classes:
                perturbed_stream = apply_perturbation(stream, fc,

mode="convex_uniform_mix")
                tau_rec = measure_regime_recovery_time(perturbed_stream,
stream)
                h_pred  = measure_predictive_horizon(perturbed_stream,
stream)
                persistence_per_feature[fc] = composite(tau_rec, h_pred)
            R1_statistic = compute_high_w_persistence_separation(
                domain_results, persistence_per_feature
            )

        CASE D2:
            mutation_tolerance = compute_position_substitution_rate(
                stream,
                phylogenetic_correction = pre_registered_method
            )
            R1_per_family = empty list
            FOR each family f IN stream.families:
                w_aggregate =
aggregate_weights_across_grid(domain_results[f])
                rho = spearman_rank_correlation(
                    w_aggregate,
                    inverse(mutation_tolerance[f])
                )
                append rho to R1_per_family
            R1_statistic = aggregate_across_families(R1_per_family)

        CASE D3:
            persistence_score = compute_post_statement_volatility_composite(
                stream,
                weights = pre_registered_composite_weights,
                window = 10_trading_days
            )
            terciles = partition_by_persistence(stream, persistence_score)
            high_w_dist  = mean_w_distribution(domain_results, terciles.high)
            low_w_dist   = mean_w_distribution(domain_results, terciles.low)
            R1_statistic = jensen_shannon_divergence(high_w_dist, low_w_dist)
```

```
    bootstrap_CIs = bootstrap_R1_statistic(
        R1_statistic,
        B = 1000,
        method = "non_parametric"
    )

    R1_verdict = {
        statistic = R1_statistic,
        cohens_d  = effect_size(R1_statistic),
        CI_lower  = bootstrap_CIs.lower_95,
        p_value   = adjusted_p_value_bonferroni(R1_statistic),
        passes    = (cohens_d > threshold_spec.d AND
                     bootstrap_CIs.lower_95 > threshold_spec.CI_lower AND
                     p_value < threshold_spec.alpha)
    }
    RETURN R1_verdict
```

A.5 R3 evaluation — intervention asymmetry

R3 paired-intervention specifications differ by domain per Sections 5.5, 6.5, and 7.4.

```
FUNCTION evaluate_R3(domain_results, stream, domain, threshold_spec):
    INPUT: domain_results — w distributions across all (K, A) cells
           stream         — the domain's stream
           domain         — D1, D2, or D3
           threshold_spec — structural/interpretive ratio > 3.0, CI lower >
2.0
    OUTPUT: R3_verdict — pass/fail with statistics

    structural_w_deltas    = empty list
    interpretive_w_deltas  = empty list

    SWITCH domain:
        CASE D1:
            structural_perturbations = {
                "transition_matrix_modification",
                "drift_rate_modification"
            }
            interpretive_perturbations = {
                "alphabet_permutation_dynamics_preserving"
            }

        CASE D2:
            structural_perturbations = {
                "relaxed_selection_simulation"
            }
            interpretive_perturbations = {
                "BLOSUM_equivalent_relabeling"
            }

        CASE D3:
            structural_perturbations = pre_registered_breakpoints  # 5
breakpoints
            interpretive_perturbations = pre_registered_language_shifts
```

```
    FOR each perturbation_set, target_list IN [
        (structural_perturbations,   structural_w_deltas),
        (interpretive_perturbations, interpretive_w_deltas)
    ]:
        FOR each perturbation p IN perturbation_set:
            stream_perturbed = apply_intervention(stream, p, domain)
            FOR each (K, A) IN cells(domain_results):
                w_baseline   = domain_results[(K, A)]
                w_perturbed  = induce_weights(stream_perturbed, K, A)
                delta = L1_norm(w_perturbed - w_baseline)
                append delta to target_list

    structural_norm    = aggregate(structural_w_deltas)
    interpretive_norm  = aggregate(interpretive_w_deltas)
    R3_ratio = structural_norm / interpretive_norm

    bootstrap_CIs = bootstrap_R3_ratio(
        structural_w_deltas,
        interpretive_w_deltas,
        B = 1000,
        method = "non_parametric"
    )

    R3_verdict = {
        ratio    = R3_ratio,
        CI_lower = bootstrap_CIs.lower_95,
        passes   = (R3_ratio > threshold_spec.ratio AND
                    bootstrap_CIs.lower_95 > threshold_spec.CI_lower)
    }
    RETURN R3_verdict
```

The intervention pairs are committed at pre-registration. D_3's structural breakpoints are the five named in Section 7.4; D_3's interpretive shifts are documented in the lexicon published with the pre-registration. D_2's relaxed-selection method is committed at pre-registration. D_1's perturbation parameters are committed at pre-registration as part of the parameter calibration locked under Section 5.4.

A.6 M5 evaluation — cross-domain admissibility gate

The M5 gate is sequenced before the outcome matrix per Section 8.2. Its computation reads vertically across domains under matched estimator philosophy, in contrast to R2 which reads horizontally within a domain.

```
FUNCTION evaluate_M5(result_store, partition, threshold_spec):
    INPUT: result_store — w distributions across {K × A × Domain}
           partition    — feature-class partitions per domain
                          (coherence-bearing vs noise)
           threshold_spec — factor-of-2 across domain pairs, 3-of-5
estimators
```

```
OUTPUT: M5_verdict — gate pass/fail with statistics

domain_set = {D1, D2, D3}
estimator_set = {K1, K2, K3, K4, K5}
operator_set = {A1, A2, A3}

# For D3, M5 reads only Level 1 (token) per Section 7.2
selected_d3_level = "Level1_token"

estimators_passing_consistency = empty list

FOR each estimator K IN estimator_set:
    ratios_per_domain = empty dictionary

    FOR each domain d IN domain_set:
        ratios_per_operator = empty list
        FOR each operator A IN operator_set:
            IF d == D3:
                w_dKA = result_store[d][selected_d3_level][K][A]
            ELSE:
                w_dKA = result_store[d][K][A]

            w_normalized = rank_normalize(w_dKA)

            coherence_features = partition[d].coherence_bearing
            noise_features     = partition[d].noise

            mean_w_coherence = mean([w_normalized[x]
                                     FOR x IN coherence_features])
            mean_w_noise     = mean([w_normalized[x]
                                     FOR x IN noise_features])
            ratio = mean_w_coherence / mean_w_noise
            append ratio to ratios_per_operator

        ratios_per_domain[d] = aggregate(ratios_per_operator)

    # Cross-domain consistency check for this estimator
    max_ratio = max(values(ratios_per_domain))
    min_ratio = min(values(ratios_per_domain))
    consistency_factor = max_ratio / min_ratio

    IF consistency_factor <= 2.0:
        append K to estimators_passing_consistency

bootstrap_CIs = bootstrap_M5_consistency(
    result_store,
    partition,
    B = 1000,
    method = "matched_resample_per_domain"
)

M5_verdict = {
    estimators_passing = estimators_passing_consistency,
    count_passing      = length(estimators_passing_consistency),
    CI_lower           = bootstrap_CIs.lower_95,
    passes             = (count_passing >= 3 AND
```

```
                                bootstrap_CIs.lower_95 >=
threshold_spec.CI_threshold)
    }
    RETURN M5_verdict
```

The function returns both the gate verdict and the per-estimator consistency factors. Under M5-fail, the per-estimator and per-domain pattern feeds the M5-fail interpretation matrix in Section 8.4: which domains exhibit R-signature pass under which estimators determines whether the outcome reads as D_1-only success, D_1-and-D_2 success, or universal-pass-under-pluralism.

A.7 Outcome matrix binding

The matrix bind sequences M5 first, then reads R-signatures conditionally per Sections 8.3 and 8.4.

```
FUNCTION bind_outcomes_to_matrix(M5_result, signature_results, matrix_spec):
    INPUT: M5_result          — output of evaluate_M5
           signature_results  — {R1_d, R2_d, R3_d} per domain
           matrix_spec        — pre-registered eight-cell consequence mapping
    OUTPUT: framework_consequence — bound interpretation

    # Aggregate within-domain signatures across domains for the matrix lookup
    R1_global = aggregate_signature_across_domains(
        [signature_results[d].R1 FOR d IN domain_set]
    )
    R2_global = aggregate_signature_across_domains(
        [signature_results[d].R2 FOR d IN domain_set]
    )
    R3_global = aggregate_signature_across_domains(
        [signature_results[d].R3 FOR d IN domain_set]
    )

    triple = (R1_global.passes, R2_global.passes, R3_global.passes)

    IF M5_result.passes:
        consequence = matrix_spec.under_M5_pass[triple]
    ELSE:
        # M5-fail interpretation reads per-domain pattern
        per_domain_pass_pattern =
compute_per_domain_pattern(signature_results)
        consequence = matrix_spec.under_M5_fail[per_domain_pass_pattern]

    flag_structurally_interesting = (
        triple == (False, True, False)   OR  # R1✗ R2✓ R3✗
        triple == (False, True, True)    OR  # R1✗ R2✓ R3✓
        (NOT M5_result.passes AND
         all_domains_pass(per_domain_pass_pattern))   # M5✗ universal R-pass
    )

    RETURN {
        triple        = triple,
```

```
        M5_pass       = M5_result.passes,
        consequence   = consequence,
        structurally_interesting = flag_structurally_interesting
    }
```

The structurally-interesting flag is computed automatically per Section 8.5 — the three configurations that exemplify failure modes the framework had previously named are tagged at the report level, ensuring no result that registers as such an exemplification can be misread as a routine outcome.

A.8 Joint distribution reporting

```
FUNCTION report_joint_distributions(result_store, signature_results,
M5_result):
    cross_signature_correlations = compute_pairwise_correlations(
        signature_results.flatten_across_grid(),
        pairs = [(R1, R2), (R1, R3), (R2, R3), (R1, M5), (R2, M5), (R3, M5)]
    )

    correlation_stability_across_grid = compute_correlation_stability(
        cross_signature_correlations,
        across = ["estimator", "operator", "domain"]
    )

    meta_coherence_finding = (
        correlation_stability_across_grid.variance < pre_registered_threshold
    )

    RETURN {
        joint_correlations          = cross_signature_correlations,
        correlation_stability       = correlation_stability_across_grid,
        meta_coherence_signal       = meta_coherence_finding,
        full_grid_w_distributions   = result_store
    }
```

A.9 Computational cost and parallelization

The grid contains 5 estimators × 3 operators × 3 domains = 45 (K, A, D) cells, plus the D_3 coarse-graining sub-experiment which adds 30 cells (2 additional levels × 15 (K, A) cells), plus the intervention reruns which multiply each cell by the perturbation count. The M5 gate evaluation re-reads the 45 primary cells without additional induction. Total induction calls in the program approximate 45 + 30 + 45 × (intervention count per domain) ≈ 200–300 induction passes.

The pseudocode is structured for embarrassing parallelism across (K, A, D) cells in Phase 1 and across perturbation reruns in R3 evaluation. K_3 (transformer prequential coding) dominates per-cell

compute; the others are inexpensive by comparison. A reasonable execution plan distributes K_3 runs across GPUs and runs K_1, K_2, K_4, K_5 on CPU. The full program completes in days, not weeks, on a small cluster, and in weeks on a single workstation with one GPU.

A.10 Reproducibility commitments.

Every random draw in the program, calibration parameter selection, Shapley coalition sampling, bootstrap resampling, perturbation random components, is seeded by the pre-registered random seed via deterministic substream derivation. Re-running the program with the same seed and the same code produces identical outputs to the precision of floating-point determinism on the target hardware. The pre-registration package includes the seed, the random-stream-derivation specification, and the version-locked dependency manifest. Independent replication is supported by direct re-execution; cross-implementation replication is supported by the fact that the program's outputs are determined by the seed, the specification, and the choice of implementation language only at the level of floating-point order of operations, which Section 8.7's bootstrap CI procedures absorb.

Appendix B — D_1 Generator Parameter Calibration Protocol

The D_1 generator specified in Section 5 has free parameters whose values must be locked at pre-registration. The locked values are not arbitrary; they are required to satisfy two structural conditions named in Section 5.4. First, no single property (A, B, C, D) explains more than 40% of total $\hat{C}$ variance under the K_3 baseline estimator with A_1 operator. Second, all four properties are in principle recoverable above the noise floor that n_1, n_2, n_3 collectively establish. This appendix specifies the calibration procedure that produces the locked values, the diagnostic measurements that determine whether candidate parameters satisfy the conditions, the pre-registration package that the procedure delivers, and the discipline that prevents the calibration's procedural freedom from contaminating the pre-registered analysis.

B.1 What is calibrated and what is not

Calibration locks the generator's structural parameters. Calibration does not lock the analysis pipeline, the estimator implementations, the operator implementations, the threshold values, the perturbation magnitudes used in R1 evaluation, or the intervention specifications used in R3 evaluation. Those are committed at pre-registration directly; they are not products of the calibration procedure.

The parameters subject to calibration are listed in B.2. Each is given a candidate range, a calibration objective, and a measurement procedure. The calibration grid sweeps the candidate ranges; the grid's outputs are the diagnostic measurements; the diagnostic measurements determine which candidate values satisfy the structural conditions; the satisfying values are locked and registered.

Calibration is performed on synthetic data generated by the D_1 generator itself, run at proposed parameter settings. No calibration draws on D_2 or D_3 data. The calibration is therefore self-contained within D_1's transparency tier.

B.2 Calibrated parameters and candidate ranges

The parameters subject to calibration, with their candidate ranges and the structural rationale for each range, are as follows.

Regime sojourn distribution parameters: The negative binomial duration distribution has mean μ_Z and dispersion parameter ψ_Z. The candidate range for μ_Z is {150, 200, 250, 300} steps. The candidate range for ψ_Z corresponds to coefficient of variation in {0.3, 0.4, 0.5}. Lower μ_Z values produce more frequent regime transitions, increasing the regime-inference signal at the cost of within-regime

stability for long-range coupling detection. Higher μ_Z values produce more stable regimes at the cost of regime-transition signal. The 200-step center of the range balances these, given the 50-step lag of the long-range coupling.

Regime transition matrix off-diagonal mass: The off-diagonal mass parameter δ_Z controls transition sparsity. The candidate range is $\delta_Z \in \{0.6, 0.7, 0.8, 0.9\}$, where the transition matrix has diagonal mass $(1 - \delta_Z)$ within a sojourn-completion event. Higher δ_Z increases regime-transition variability; lower δ_Z makes regime transitions more deterministic. Three regime states means each off-diagonal cell receives δ_Z / 2 mass under the equiprobable-transition design.

Regime-conditional emission distance for f_A: The pairwise total-variation distance between the three regime-conditional categorical distributions for f_A, denoted d_A. Candidate range: $d_A \in \{0.3, 0.4, 0.5, 0.6\}$. Higher d_A makes regime detection from f_A trivial; lower d_A buries f_A's signal in the noise floor. The calibration objective is to find d_A values that make Property A clearly recoverable but not so saturated that it dominates $\hat{\mathrm{C}}$ variance.

Long-range coupling lag τ: The lag at which $f_{B(t)} = g(X_{\{t-\tau\}})$. Candidate range: $\tau \in \{30, 50, 70\}$. The 50-step center matches the body of the paper. Lower τ values push f_B's coupling toward local detection, where dictionary compression may recover it; higher τ values push it beyond the practical context window of any estimator in the class. The objective is to land f_B in the regime where state-space and predictive-neural estimators recover it but dictionary and short-context n-gram do not.

Coalitional pair encoding strength: The mutual information between (f_{C_1}, f_{C_2}) jointly and (Z_t, s_t) jointly, denoted I_C. Candidate range: $I_C \in \{0.5, 0.7, 0.9, 1.1\}$ bits. Lower I_C makes the coalition undetectable even with Shapley-style operators; higher I_C makes the marginal signal in f_{C_1} alone or f_{C_2} alone large enough that LOO partially recovers the structure, defeating the operator-decoupling design.

Drift carrier rate: The drift rate parameter ρ_D for f_D's slow categorical drift, expressed as the within-regime drift accumulation as a fraction of the regime-conditional emission distance. Candidate range: $\rho_D \in \{0.1, 0.2, 0.3\}$ per regime sojourn. Lower ρ_D makes f_D indistinguishable from stable noise; higher ρ_D makes f_D's drift overwhelm the regime-conditional structure.

Distractor noise calibration parameters: The Zipf exponent α_{Zipf} for n_2 and the dominant-symbol mass parameter $\gamma_{\mathrm{lowfreq}}$ for n_3. Candidate ranges: $\alpha_{\mathrm{Zipf}} \in \{1.2, 1.5, 1.8\}$ and $\gamma_{\mathrm{lowfreq}} \in \{0.5, 0.6, 0.7\}$. The calibration objective is to give n_2 and n_3 surface statistical structure that frequency-based estimators may register as coherence-bearing, providing the false-positive challenge the design requires.

Stream length and replicate count: Per-stream length N is fixed at 50,000 steps as committed in Section 5.1; replicate count M is fixed at 20. These are not subject to calibration. They are committed at pre-registration as analysis-design parameters.

The full candidate grid spans 4 × 3 × 4 × 4 × 3 × 4 × 3 × 3 × 3 = 31,104 parameter combinations. Exhaustive search is computationally feasible at moderate cost given the reduced calibration stream length specified in B.3.

B.3 Calibration stream specification

The calibration runs on streams of reduced length N_{cal} = 10,000 steps and reduced replicate count M_{cal} = 5. The reduction is solely a computational concession; the analytic stream length and replicate count revert to the full N = 50,000 and M = 20 once parameters are locked. Reduced-length calibration produces noisier diagnostic measurements but does not change which parameter regions satisfy the structural conditions, because the conditions are defined as variance shares and recoverability margins that scale predictably with stream length.

For each candidate parameter combination in the grid, the calibration generates 5 streams of 10,000 steps each, runs the K_3 × A_1 baseline through the induction pipeline, and records the diagnostic measurements specified in B.4.

B.4 Diagnostic measurements per candidate combination

Each candidate combination yields a diagnostic record containing the following measurements.

Per-property variance share: For each property class (A: $\{f_A\}$, B: $\{f_B\}$, C: $\{f_{C_1}, f_{C_2}\}$, D: $\{f_D\}$, noise: $\{n_1, n_2, n_3\}$), the share of total $\hat{C}$ variance attributable to ablating that property class is computed via group-ablation contribution. The shares sum to 1 by construction over the five classes (A, B, C, D, noise). The calibration condition is that no single non-noise property class share exceeds 0.40.

Per-property recoverability margin: For each property class, the recoverability margin is the difference between the property class's mean induced w (under K_3 × A_1 baseline) and the noise class's mean induced w, normalized by the standard deviation of induced w over the noise class. Margins below 1.0 register as poor recoverability; margins above 3.0 register as saturated recoverability. The calibration condition is that all four property class margins fall in [1.0, 3.0].

Cross-replicate stability: The Spearman rank correlation of induced w across the 5 calibration replicates, averaged pairwise. Below-threshold cross-replicate stability indicates that the

parameter combination is too noise-dominated to support reliable analysis at full stream length. The calibration condition is that mean cross-replicate Spearman exceeds 0.7.

Estimator-dependence diagnostic: A reduced run on a single replicate using $K_1 \times A_1$ and $K_4 \times A_1$ in addition to $K_3 \times A_1$, computing per-property recoverability margins under each. The calibration condition for Property B is that K_4's recoverability margin exceeds K_1's by at least 1.0 in standard deviation units, ensuring that f_B's coupling is genuinely lag-dependent in the regime where state-space estimators outperform dictionary estimators. The calibration condition for Property C is that $A_2 \times K_3$ on a single test run produces a coalition recoverability margin exceeding $A_1 \times K_3$'s by at least 1.0 in standard deviation units, ensuring that f_C is genuinely operator-dependent.

A candidate combination passes calibration if and only if it satisfies all four conditions: per-property variance shares below 0.40, per-property recoverability margins in [1.0, 3.0], cross-replicate Spearman above 0.7, and estimator-dependence and operator-dependence margins above 1.0 standard deviation unit.

B.5 Selection rule among passing combinations

The candidate grid will typically yield multiple passing combinations. The selection rule among them is committed in advance and is not subject to discretion at calibration time.

The selected combination is the one that maximizes the minimum recoverability margin across the four properties. The criterion privileges parameter combinations that make the diagnostic test as resolved as possible. The property with the smallest margin is the property the program will have the hardest time analyzing reliably, and selecting the combination that maximizes that smallest margin produces the most diagnostically useful generator.

If multiple combinations tie at the maximin margin within numerical precision, the tiebreaker is minimum total variance share concentrated in any one property class, again maximizing the diagnostic uniformity of the design. If multiple combinations tie on both criteria, the tiebreaker is parameter-set centrality, the combination closest to the geometric center of the candidate grid in normalized coordinates. Centrality serves as a robustness heuristic, parameter combinations near grid edges are more sensitive to small misspecifications of the candidate ranges, while central combinations have more candidate neighbors that also pass.

The selection rule produces a unique winner from any non-empty passing set with probability one over numerical precision, eliminating selection discretion at calibration close.

B.6 The pre-registration package

When the calibration procedure terminates, the pre-registration package is assembled and registered. It contains the following items.

The locked parameter values for all calibrated parameters: $\mu_Z, \psi_Z, \delta_Z, d_A, \tau, I_C, \rho_D, \alpha_{\text{Zipf}}, \gamma_{\text{lowfreq}}$.

The fixed analysis-design parameters: N = 50,000, M = 20, alphabet sizes for each feature, transition matrix structure, sojourn distribution family.

The diagnostic measurements at the locked parameter values; per-property variance shares, per-property recoverability margins, cross-replicate stability, and estimator/operator dependence diagnostics. These serve as a reproducibility check; an independent re-execution of the calibration with the registered seed should recover them exactly.

The full calibration grid output: the diagnostic record for every candidate combination evaluated during calibration. This is published as a supplementary data file. Its purpose is auditability: a reader can verify that the locked combination is indeed selected by the registered selection rule applied to the registered diagnostic measurements.

The random seed used for calibration stream generation, the substream derivation specification, and the software version manifest. These permit deterministic re-execution of the calibration itself, not just the analysis that uses its outputs.

The selection rule, in the exact form specified in B.5, included verbatim in the pre-registration text. Including the rule in the registration prevents post-hoc reinterpretation of which combination "should have" been selected.

A statement that the calibration procedure has terminated and that no further parameter adjustment is permitted under the pre-registration. Subsequent discoveries, that a different parameter combination would have produced cleaner results, that the candidate ranges were misspecified, that the diagnostic conditions were too lax or too strict, do not authorize re-calibration. They authorize a new pre-registration for a new analysis, separately registered and reported as such.

B.7 The discipline that separates calibration freedom from analysis discipline

Calibration is procedurally free in a specific sense; the candidate ranges, the structural conditions, the selection rule, and the diagnostic measurements are all designed with full knowledge of the analysis they will support. The freedom is bounded by the requirement that all of these be committed before the calibration runs. Once the candidate ranges are set, the conditions are written down, the

selection rule is fixed, and the diagnostic measurement procedures are specified, the calibration becomes a deterministic function of the random seed.

The discipline that separates calibration freedom from analysis discipline operates at three points.

The first point is the freezing boundary. The candidate ranges, structural conditions, selection rule, and measurement procedures are fixed before the calibration generates any data. After that boundary, no element of the calibration design is adjustable. If the calibration grid contains no passing combinations, the analysis does not proceed; the program is reconceived and a new pre-registration is filed. The pre-registration cannot be amended in response to calibration failure to make calibration succeed.

The second point is the registration boundary. After calibration terminates and the locked combination is selected, the pre-registration package is assembled and registered with a third-party service that timestamps and archives the package. The service used must support cryptographic verification of the registration timestamp. The registration occurs before any analysis runs. The package, once registered, cannot be modified.

The third point is the analysis boundary. Once registered, the pre-registration governs the analysis as committed. Departures from the registered analysis design, using a different selection rule retroactively, applying different statistical thresholds, swapping intervention specifications, are excluded. Departures necessitated by discovery during analysis (a software bug in the implementation, a data-loading error, a hardware failure that corrupts a replicate) require explicit documentation in the analysis report, with a precise description of what changed and why. The default rule is that the analysis report is bound to the pre-registration; deviations are reported as deviations, not absorbed into the analysis as if they had been planned.

The three boundaries together convert what could otherwise be unlimited researcher discretion at the calibration phase into a procedure whose outputs are determined by the registered seed and the registered design. Calibration freedom is freedom to specify the program's structural parameters before the program runs. It is not freedom to adjust those parameters once the program is running, and it is not freedom to renegotiate them once results are visible.

B.8 Calibration's epistemic status under Synthetic Epistemology

The calibration procedure does not violate the framework's foreclosure of correspondence-truth. It does not appeal to ground truth about the substrate; D_1's substrate is written, not discovered. Calibration calibrates the relationship between the substrate's structure (which the writer specifies)

and the diagnostic measurements that read out of the substrate through the $K_3 \times A_1$ baseline. The procedure ensures that the written substrate exposes the four-property design at the resolution the diagnostic deliverable in Section 5.7 requires. It does not validate the framework's apparatus against an external standard.

The procedure's epistemic status is internal to D_1's transparency tier. It is the writer's procedure for ensuring that what they wrote is readable by their instruments at the resolution they require. The framework's universality claim, evaluated against the locked generator, is untouched by the calibration: the question of whether induced w recovers the substrate's structure under the framework's apparatus is not what calibration answers. Calibration answers the prior question of whether the substrate's structure is in principle recoverable above the noise floor, which is a precondition for the framework's apparatus to be tested at all.

Reframing calibration locks the test surface; the test surface is then run against the apparatus. The two operations are sequenced, separated by the pre-registration boundary, and committed to distinct epistemic functions. The calibration is internal to the substrate. The analysis is internal to the framework. Neither is internal to anything outside both.

Appendix C — Statistical Analysis Plan

This appendix specifies the statistical analysis plan in formal terms. The plan binds the bootstrap procedures used for confidence-interval construction across all four conditions, the multiple-comparison adjustment applied across the program's full grid, the joint-distribution reporting templates that operationalize Section 8.6's commitment to reporting beyond marginal pass/fail, the M5 gate evaluation procedure as a self-contained statistical operation, and the sequencing of M5 evaluation before the eight-cell outcome matrix as a procedural requirement that conditions all downstream interpretation.

C.1 Test sequencing

The analysis proceeds in a fixed sequence that is not adjustable post-hoc. The sequence is committed at pre-registration alongside the locked thresholds in Section 8.7 and the pre-registration package specified in Appendix B.6.

Step 1: Phase 1 of the program runs the full {K × A × Domain} grid as specified in Appendix A, producing 45 induced-w distributions plus 30 distributions for the D_3 coarse-graining sub-experiment.

Step 2: M5 admissibility evaluation runs on the Phase 1 output, as specified in C.5. The M5 verdict is computed and recorded.

Step 3: Within-domain R-signature evaluations (R1, R2, R3 per domain) run on the Phase 1 output, as specified in C.2 through C.4. Bootstrap CIs are constructed and pass/fail verdicts are recorded per domain and aggregated globally.

Step 4: Multiple-comparison adjustment is applied across the full set of tests run in Steps 2 and 3. The adjustment is described in C.6.

Step 5: Outcome matrix interpretation reads the M5 verdict first; under M5-pass, the eight-cell matrix in Section 8.3 is consulted; under M5-fail, the M5-fail interpretation matrix in Section 8.4 is consulted. The sequencing is non-negotiable: the M5 verdict conditions the matrix's reading, and the matrix is not interpreted independently of the M5 verdict.

Step 6: Joint-distribution reporting per C.7 is computed and reported. The joint distributions are reported regardless of the marginal verdicts, because the joint structure is itself informative independently of whether individual conditions pass.

Step 7: The full analysis report is assembled with the verdict at each layer, the bootstrap statistics, the joint distributions, the structurally-interesting flag where applicable, and any deviations from the pre-registered analysis design with explicit documentation per the discipline in Appendix B.7.

C.2 Bootstrap procedure for R2

R2 is the cross-philosophy convergence test, evaluated within each domain via pairwise Spearman rank correlation across the 5×3 grid (Section 3.3, Appendix A.3).

For each domain d, let G_d denote the set of induced-w distributions across the 15 (K, A) cells in that domain. Let P_{cross} denote the set of cell pairs $(\text{cell}_i, \text{cell}_j)$ where cell_i and cell_j belong to philosophically-decoupled estimator classes, with the philosophical classes specified in Section 3.1 (coding-scheme class: K_1, K_2, K_3; state-space class: K_4; algorithmic class: K_5).

The R2 statistic for domain d is the median Spearman rank correlation across P_{cross}.

Bootstrap procedure:

```
FUNCTION bootstrap_R2(domain_d, B = 1000, block_method):
    medians = empty list
    FOR b FROM 1 TO B:
        IF block_method == "time_window_resample":         # D1, D3
            block_length = mean_regime_sojourn (D1) OR
                           one_statement_unit (D3)
            resampled_stream = circular_block_bootstrap(
                domain_d.stream, block_length
            )
        ELIF block_method == "ortholog_resample":          # D2
            resampled_stream = ortholog_resample(
                domain_d.stream, with_replacement = TRUE
            )

        rerun the induce_weights pipeline on resampled_stream
        for all 15 (K, A) cells, producing G_d^(b)

        compute pairwise Spearman across P_cross on G_d^(b)
        median_b = median of those correlations
        append median_b to medians

    sort medians ascending
    CI_95_lower = medians[25]    # 2.5th percentile
    CI_95_upper = medians[975]   # 97.5th percentile
    point_estimate = median(domain_d.cross_philosophy_spearman)

    RETURN {
        point_estimate,
        CI_95_lower,
        CI_95_upper,
        bootstrap_distribution = medians
    }
```

The resampling unit is the structural temporal unit for D_1 and D_3 (one regime sojourn for D_1, one statement-unit for D_3) and the ortholog for D_2. Resampling at the structural unit preserves within-unit dependence structure that resampling at finer granularity (individual time-steps, individual residues) would destroy. Block length for D_1 is set equal to the locked μ_Z from Appendix B; the choice is fixed at pre-registration once calibration completes.

R2 passes for domain d if the bootstrap point estimate exceeds 0.6 and the bootstrap lower 95% CI exceeds 0.4. Global R2 aggregates per-domain verdicts: R2 passes globally if it passes in all three domains under the multiple-comparison adjustment in C.6.

C.3 Bootstrap procedure for R1

R1 is the persistence prediction test, with operationalization differing by domain (Sections 5.5, 6.3, 7.3, Appendix A.4).

For D_1, the R1 statistic is Cohen's d between high-w-feature persistence outcomes and low-w-feature persistence outcomes, where high/low partitioning uses the median induced-w per cell. For D_2, the R1 statistic is the Spearman rank correlation between aggregated induced w and inverse mutation tolerance per family, aggregated across families. For D_3, the R1 statistic is the Jensen-Shannon divergence between mean-w distributions in high-persistence and low-persistence statement terciles.

Bootstrap procedure (non-parametric, resampling unit per domain):

```
FUNCTION bootstrap_R1(domain_d, B = 1000):
    statistics = empty list
    FOR b FROM 1 TO B:
        IF domain_d == D1:
            resampled = bootstrap_replicate_streams(domain_d.replicates)
        ELIF domain_d == D2:
            resampled = bootstrap_families(domain_d.families)
        ELIF domain_d == D3:
            resampled = bootstrap_statements(domain_d.corpus)

        compute R1_statistic on resampled
        compute Cohen's d (D1, D3) or family-aggregate Spearman (D2)
        append to statistics

    sort statistics ascending
    CI_95_lower = statistics[25]
    CI_95_upper = statistics[975]
    point_estimate = R1_statistic on full data
    p_value = bootstrap_p_value(statistics, null_hypothesis = no_separation)

    RETURN {
        point_estimate,
        cohens_d_or_equivalent,
```

```
        CI_95_lower,
        CI_95_upper,
        p_value,
        bootstrap_distribution = statistics
    }
```

The resampling units are: for D_1, the 20 replicate streams (resample with replacement to produce 20 bootstrap streams per replicate); for D_2, the families (resample with replacement to produce a same-size family set); for D_3, the statements (resample with replacement to produce a same-size statement set, preserving within-statement structure but resampling across statements).

R1 passes for domain d if Cohen's d (or domain-equivalent statistic standardized to d-units) exceeds 0.5, the bootstrap lower 95% CI exceeds 0.3, and the bootstrap-derived p-value is below 0.01 after multiple-comparison adjustment per C.6. R1 passes globally if it passes in all three domains.

C.4 Bootstrap procedure for R3

R3 is the intervention asymmetry test (Sections 5.5, 6.5, 7.4, Appendix A.5).

For each domain, the R3 ratio is structural $\|\Delta w\|$ aggregated across structural perturbations divided by interpretive $\|\Delta w\|$ aggregated across interpretive perturbations.

Bootstrap procedure (non-parametric, perturbation-pair resampling):

```
FUNCTION bootstrap_R3(domain_d, B = 1000):
    ratios = empty list
    structural_deltas    = compute_structural_w_deltas(domain_d)
    interpretive_deltas  = compute_interpretive_w_deltas(domain_d)

    FOR b FROM 1 TO B:
        struct_resampled = bootstrap_with_replacement(structural_deltas)
        interp_resampled = bootstrap_with_replacement(interpretive_deltas)

        ratio_b = aggregate(struct_resampled) / aggregate(interp_resampled)
        append ratio_b to ratios

    sort ratios ascending
    CI_95_lower = ratios[25]
    CI_95_upper = ratios[975]
    point_estimate = aggregate(structural_deltas) /
aggregate(interpretive_deltas)

    RETURN {
        point_estimate,
        CI_95_lower,
        CI_95_upper,
        bootstrap_distribution = ratios
    }
```

The aggregation function is committed at pre-registration and is the geometric mean across perturbations within each class. Geometric mean is chosen over arithmetic mean to prevent any single high-magnitude perturbation from dominating the aggregate, which would distort the structural-to-interpretive ratio.

The bootstrap resamples perturbation events independently within each class, preserving the structural-vs-interpretive class distinction at the resampling level. Cross-class resampling is excluded by design; the ratio is between two distinct populations that the test is constructed to keep separate.

R3 passes for domain d if the point estimate ratio exceeds 3.0 and the bootstrap lower 95% CI exceeds 2.0. R3 passes globally if it passes in all three domains under multiple-comparison adjustment per C.6.

C.5 M5 gate evaluation procedure

M5 is the cross-domain admissibility gate (Section 2.5, Section 8.2, Appendix A.6). Its evaluation is sequenced before the eight-cell outcome matrix and conditions all downstream interpretation.

For each estimator K_i, the M5 procedure computes:

For each domain $d \in \{D_1, D_2, D_3\}$ and each operator A_j:

1. Retrieve the induced-w distribution $w_{\{d,K_i,A_j\}}$.
2. Apply rank normalization within the domain: replace each w-value with its rank divided by the number of features, producing rank-normalized w in [0, 1].
3. Identify coherence-bearing features and noise features per the partition committed in Section 5.6 (D_1), Section 6.6 (D_2), or Section 7.6 (D_3).
4. Compute mean rank-normalized w over coherence-bearing features (μ_C) and over noise features (μ_N).
5. Compute the ratio $r_{\{d,K_i,A_j\}} = \frac{\mu_C}{\mu_N}$.

Aggregate across operators within (estimator, domain) to produce r_v = geometric mean over A_j.

For each estimator K_i, compute the cross-domain consistency factor:

$$f_{\{K_i\}} = \frac{\max_d r_{\{d,K_i\}}}{\min_d r_{\{d,K_i\}}}$$

K_i passes the per-estimator consistency check if $f_{\{K_i\}} \leq 2.0$.

The M5 gate passes if at least 3 of the 5 estimators pass the per-estimator consistency check.

Bootstrap procedure for M5:

```
FUNCTION bootstrap_M5(result_store, partition, B = 1000):
    consistency_factors_distribution = empty matrix [5 estimators × B]
    pass_count_distribution = empty list

    FOR b FROM 1 TO B:
        # Joint resampling: each domain resamples independently in its
```

```
        # natural unit, but the resampling is matched across estimators
        # to preserve covariance structure
        D1_resample = block_resample(D1.stream, block_length = mean_sojourn)
        D2_resample = ortholog_resample(D2.stream)
        D3_resample = statement_resample(D3.corpus)

        # Re-induce w across the grid for resampled streams
        w_resampled = induce_full_grid(D1_resample, D2_resample, D3_resample)

        FOR each estimator K_i:
            FOR each domain d:
                ratio_b_{d, K_i} = compute_normalized_ratio(
                    w_resampled[d][K_i],
                    partition[d]
                )
            f_b_{K_i} = max(ratio_b) / min(ratio_b)
            consistency_factors_distribution[K_i, b] = f_b_{K_i}

        pass_count_b = count of K_i where f_b_{K_i} <= 2.0
        append pass_count_b to pass_count_distribution

    # M5 passes if at least 3 of 5 estimators pass with adequate CI
    pass_count_lower_CI = percentile(pass_count_distribution, 2.5)

    RETURN {
        per_estimator_factors = consistency_factors_distribution,
        per_estimator_passes = which K_i achieve f_K_i <= 2.0 in point
estimate,
        pass_count_point = count of K_i passing in point estimate,
        pass_count_lower_CI,
        gate_passes = (pass_count_point >= 3 AND pass_count_lower_CI >= 3)
    }
```

The bootstrap requires that the lower 95% CI on the pass count itself reach 3 of 5, not just the point estimate. This ensures the gate verdict is robust to within-domain sampling variability.

The M5 verdict feeds the outcome matrix sequencing in C.1 Step 5: M5-pass licenses the eight-cell matrix in Section 8.3; M5-fail directs interpretation to the M5-fail matrix in Section 8.4 with the per-domain R-signature pattern as the additional input.

C.6 Multiple-comparison adjustment

The program runs a large number of statistical tests across the {K × A × Domain} grid. Without adjustment, the family-wise error rate would be inflated, and individual cell-level passes could be confounded with chance rejection of nulls. The adjustment is committed at pre-registration and applied uniformly across the program's tests.

The total number of primary tests subject to adjustment is:

- 45 cells in the primary {K × A × Domain} grid (5 estimators × 3 operators × 3 domains)

- 30 cells in the D_3 coarse-graining sub-experiment (5 estimators × 3 operators × 2 additional levels: Level 2 and Level 3; Level 1 is included in the primary grid)
- The M5 cross-domain comparisons evaluated per estimator (5 comparisons)
- The aggregate R1, R2, R3 verdicts per domain (9 verdicts: 3 conditions × 3 domains)

Total: 89 primary tests.

The Bonferroni adjustment applies to the unadjusted p-values and to the bootstrap CI threshold tests. Under Bonferroni at family-wise α = 0.05, the per-test α becomes 0.05 / 89 ≈ 0.000562. Equivalently, the bootstrap CI threshold for individual test passes is computed at the (1 − 0.05/89)-confidence level rather than the 95th-percentile level for the family-wise error guarantee.

For the locked thresholds in Section 8.7 (R1: $p < 0.01$, lower CI > 0.3; R2: lower CI > 0.4; R3: lower CI > 2.0; M5: pass-count lower CI ≥ 3), the Bonferroni-adjusted forms are:

- R1: per-test $p < 0.01 / 89 \approx 1.12 \times 10^{-4}$; lower CI at the (1 − 0.05/89) confidence level
- R2: lower CI at the (1 − 0.05/89) confidence level on the median Spearman
- R3: lower CI at the (1 − 0.05/89) confidence level on the structural-to-interpretive ratio
- M5: lower CI at the (1 − 0.05/89) confidence level on the pass-count

The adjustment is conservative. Bonferroni assumes worst-case dependence among tests; if tests are independent or positively correlated, the adjustment over-corrects and the program will under-reject. The conservative posture is appropriate for a framework that already faces the challenge of unfalsifiable universality. The adjustment ensures that pass verdicts cannot be attributed to multiple-testing inflation, and the cost of over-correction is borne by the framework rather than by its critics.

A secondary report uses Holm-Bonferroni step-down adjustment as a less-conservative comparison, with the secondary report serving as a sensitivity check rather than as the primary verdict-producing analysis. Discrepancies between Bonferroni and Holm-Bonferroni verdicts are reported as informative findings about the framework's sensitivity to adjustment method choice.

C.7 Joint distribution reporting templates

Section 8.6 commits to reporting joint distributions across the four conditions, not just marginal pass/fail verdicts. This appendix specifies the reporting templates.

Template 1: Pairwise condition correlations across the {K × A × Domain} grid.

For each pair of conditions (R1, R2), (R1, R3), (R2, R3), (M5, R1), (M5, R2), (M5, R3), report:

- The Spearman rank correlation across the grid cells where both conditions yield a per-cell statistic
- The bootstrap 95% CI on that correlation
- A scatterplot of one condition's per-cell statistic against the other's

The pairwise correlations characterize whether the conditions move together across the grid or are independent. Strong correlation between R1 and R3, for example, would indicate that domains where induced w predicts persistence are the same domains where it discriminates intervention class; weak correlation would indicate the two conditions track distinct framework commitments.

Template 2: Per-cell verdict heatmap.

A 5 × 3 × 3 (K × A × Domain) cube where each cell is colored by its R1/R2/R3 verdict combination, with M5 status overlaid as a domain-level border indicator. The heatmap permits visual identification of patterns that no marginal report would expose, which estimators tend to pass which conditions, which operators tend to pass which conditions, which domains tend to pass which conditions, and how M5-pass and M5-fail conditions interact with the within-domain pattern.

Template 3: Cross-grid correlation stability metric.

The pairwise correlations from Template 1 are themselves computed within each level of the grid (within each estimator, within each operator, within each domain). The variance of the correlation across these slices is reported as the correlation stability metric. Low variance registers as a meta-coherence finding per Section 8.6; the framework's outputs covary in structurally consistent ways across the grid. High variance registers as a contingency finding; the framework's outputs covary in ways that depend on which grid slice produced them.

Template 4: Joint pass-rate report.

The $2^4 = 16$ possible joint outcomes across (M5, R1, R2, R3) per domain, and the 8-cell collapse for M5 alone with the R-triple, are reported with their observed counts and bootstrap-CI'd estimated frequencies. The report is the operational realization of the outcome matrix in Section 8. Each cell of the matrix has an observed pass/fail status, and the joint pass-rate report shows the full joint distribution rather than just the marginal verdicts the matrix consumes.

C.8 Pre-registered exclusion of post-hoc specification changes

Once the pre-registration package per Appendix B.6 is registered, no element of the statistical analysis plan in this appendix is adjustable. The bootstrap B-counts, the resampling units, the

multiple-comparison adjustment method, the joint-distribution reporting templates, the M5 gate procedure, and the test sequencing are all bound at registration and held through the full analysis.

If an unforeseen issue arises during analysis, a software bug in the bootstrap implementation, a data-loading error, a hardware failure that corrupts a replicate, or a numerical instability that prevents a specific (K, A) cell from completing, the analysis report documents the issue, describes the remediation, and explicitly distinguishes between bug-fix-driven changes (which preserve the analysis design and are documented as such) and design changes (which are excluded under the pre-registration). Design-change-driven re-analysis requires a new pre-registration filed separately and reported separately.

The discipline imported here is the same as in Appendix B.7; deviations are reported as deviations, not absorbed into the analysis as if they had been planned. The discipline operationalizes the framework's commitment, in Synthetic Epistemology's terms, that drift is recursive coherence loss. Drift in the analysis design is drift in the framework's coherence with its own pre-registered architecture, and the analysis report is the audit surface where any such drift must register explicitly.

Appendix D — Glossary of Substrate Terms

This glossary compresses the corpus's substrate vocabulary to one-paragraph definitions for readers approaching from the empirical side without full corpus exposure. The definitions are operational rather than expository; they are written for an engineer or empiricist who needs to know what each term commits the framework to, not for a philosopher who would want the full conceptual lineage. Cross-references to the body of the paper are given where the operational use is most consequential. A reader using this glossary as the entry point should follow the cross-references back into the paper rather than treating the glossary entries as standalone definitions.

Adaptive Realism. The framework's ontology. The real is defined as that which persists across recursive coherence arbitration, or that which survives, under perturbation, in the substrate's own ongoing negotiation with entropy. Reality is not a static external bedrock to which symbols correspond; it is the dynamic set of structures that maintain themselves through adaptive coherence maintenance. This commitment forecloses correspondence-truth as a category error; there is no oracle layer beneath the symbol stream from which a verdict on a claim can be drawn. The empirical consequence for this paper is that validation cannot proceed by comparing induced w to authoritative external annotations of "structural" features, because such annotations would import the static-absolute bedrock the framework eliminates. R1 (persistence prediction, Section 2.2) operationalizes Adaptive Realism's commitment; induced w is tested against the framework's own ontological criterion (persistence under perturbation) rather than against external truth.

Synthetic Epistemology. The framework's epistemology. Knowledge systems are validated through coherence rather than correspondence. A symbolic system is valid to the extent that it maintains and extends its own internal consistency across operations, with structural fidelity, regenerative capacity, and informational economy as the operational criteria. Five core principles: coherence over correspondence, recursion over reference, compression as validation, observer-pluralism as structure, and drift = ΔC loss (failure is recursive coherence loss, not deviation from truth). The empirical consequence for this paper is that validation operates through coherence-with-coherence comparisons across the program's four conditions, not through coherence-against-truth comparisons. The transparency hierarchy in Section 4 dissolves the synthetic-versus-empirical illusion by treating all three domains as compressions of generators of differing transparency, none of which has unmediated contact with a bedrock the others lack.

Symbolic lock-in. The failure mode of the symbolic layer. A symbolic system enters lock-in when its internal coherence becomes self-referentially closed and decoupled from the adaptive substrate it was built to track. The system retains logical or procedural consistency within its own framework but no longer registers feedback from the persistence-qualified structures that originally grounded it. Once lock-in occurs, the symbolic layer resists correction, amplifies confirming signals, prunes external perturbations as noise, and shifts its persistence logic from reality-tracking to self-reproduction. The corpus's central diagnostic depends on the substrate-versus-symbol distinction being detectable through structural rather than purely interpretive analysis. R3 (intervention asymmetry, Section 2.3) operationalizes the diagnostic; induced w must distinguish structural perturbation (substrate change) from interpretive perturbation (pure recoding) in order for the framework to claim it can detect symbolic lock-in empirically. The structurally-interesting failure cells in Section 8.5 (including R1✗ R2✓ R3✗) are configurations in which the framework would correctly diagnose lock-in of itself, registering the framework's own measurement operation as having drifted into self-referential closure.

Coherence (C). The framework's central metric. Coherence is recursive structural reinforcement, or the degree to which a system's structure participates in its own persistence across iterations. A coherence proxy $\hat{C}$ is an empirical estimator of C derived from observed streams; in this paper's program, $\hat{C}$ is computed via predictive-logloss form across the five estimator classes specified in Section 3.1. Coherence is not a property of static configurations; it is a property of dynamics across recursive iteration. The framework's universality claim is that coherence operates as the same kind of quantity across physical, biological, cognitive, and institutional substrates, with the same structural logic governing what persists in each. The induced-w functional in this paper attempts to operationalize that claim at the cross-domain level.

ΔC (coherence delta). The change in coherence over a specified interval or under a specified perturbation. $\Delta C > 0$ registers as adaptive success; $\Delta C < 0$ registers as drift, and persistent $\Delta C < 0$ across recursive iterations is the operational signature of symbolic lock-in or substrate failure. In the program specified here, $||\Delta w||$ is the L1 magnitude of induced-weight change under intervention and the operationalization of ΔC at the measurement-operation level for R3 evaluation. The structural-versus-interpretive intervention pair in R3 produces two ΔC measurements per perturbation type, and the ratio between them is the R3 statistic.

$\Delta^2 C$ (coherence curvature). The second-order rate of change of coherence, or how rapidly the coherence trajectory accelerates or decelerates. $\Delta^2 C$ is the quantity Recursive Arbitration (axiom 4 in the corpus's operational layer) maximizes; optimal action sequences are those that increase coherence curvature while preserving future option space. $\Delta^2 C$ is more diagnostic than ΔC for forward-looking trajectory selection, because a system with positive ΔC but flattening $\Delta^2 C$ is approaching a local coherence plateau and may need pruning to escape, while a system with positive $\Delta^2 C$ is ascending the coherence gradient with adaptive headroom. $\Delta^2 C$ does not enter the program's primary measurement layer in this paper, but it appears in the framework's ontology of arbitration and is referenced in the corpus's broader claims about adaptive trajectory selection.

Ω (Spectrum of Possibility). The set of feasible coherence-increasing trajectories from the system's current state. Ω is the space of futures the system can reach under valid action policies, bounded by the system's persistence constraints and adaptive capacity. The framework's commitment to preserving Ω, preventing premature collapse of future option space, is structural rather than aesthetic. A system that prunes Ω prematurely loses the flexibility required for adaptive response, and the volatility bound on coherence ($\sigma_C \leq \theta_{\max}$) exists to prevent runaway pruning. Structural intervention in R3 is operationalized as Ω-modifying perturbation, changing the substrate's constraint topology in ways that alter what futures are reachable. Interpretive intervention is Ω-preserving by design; pure recoding leaves the substrate's reachable-future set unchanged. The structural-vs-interpretive distinction at the R3 level is therefore the operational form of the Ω-modifying-vs-Ω-preserving distinction at the framework's ontological level.

η (responsiveness, eta). The system's responsiveness to entropy gradients, or how rapidly its adaptive logic adjusts to environmental change. η is one of the parameters in the universal coherence update law ($X_{\{t+1\}} = X_t + F(\nabla C, \nabla S, \Omega, \eta, \lambda)$) and regulates the speed at which the system tracks coherence-increasing trajectories. High η registers as rapid adaptation at the cost of stability; low η registers as stable but potentially slow tracking. η does not appear in the program's empirical measurements directly; it is a parameter of the framework's substrate-level dynamics and is referenced in the broader corpus's claims about adaptive responsiveness.

λ (leakage). The internal coherence loss rate. λ measures the rate at which the system's coherence dissipates without external perturbation, or the entropy production internal to the system's own

dynamics. The persistence equation requires C to grow faster than $\lambda \cdot C$ for the system to persist ($\frac{dC}{dt} = \Gamma \cdot C^n - \lambda \cdot C + \eta \cdot (\nabla S \cdot \nabla \Omega)$); when λ exceeds the coherence-generation rate, the system collapses regardless of external conditions. λ does not enter the program's empirical measurements directly. It is a parameter of the framework's substrate-level dynamics.

Coherence-weighted information measures. Extensions of Shannon's information-theoretic framework that assign each source symbol a weight $w(x) \in [0, 1]$ proportional to its expected contribution to recursive structural stability. Coherence entropy: $H_{w(X)} = \sum_x p(x)\, w(x)[-\log p(x)]$. Coherence mutual information: $I_w(X;\, Y)$ = coherence-weighted analog of Shannon mutual information. Coherence channel capacity: $C_C = \max_{\{p(x)\}} I_w(X;\, Y)$. Selective compression: the minimum expected code length required to preserve all coherence-bearing structure equals the coherence entropy, with achievability via weighted typical-set coding. The induced-weights program specified in this paper is the operational route by which $w(x)$ is derived from observed streams rather than hand-assigned, making the coherence-weighted measures empirically estimable rather than theoretically posited.

M1: Boundedness and monotonicity. The first admissibility commitment on the induced-w functional. Weights must be bounded in [0, 1] and monotone in a coherence-relevance score $\rho(x)$: if $\rho(x)$ increases, $w(x)$ must not decrease. M1 prevents post-hoc reassignment of relevance and bounds w against unacknowledged rescaling of probability or utility. M1 is satisfied by construction in the induction pipeline through the $\sigma(\beta\rho(x))$ mapping in Step D of the derivation pipeline.

M2: Coherence-consistency under coarse-graining. The second admissibility commitment. If micro-symbols are aggregated into macro-symbols, the induced weights on macro-symbols must be functions of micro-weights and micro-probabilities that preserve relevance ordering and do not flip under benign regrouping. M2 is not a numerical-equality requirement; it is an ordering-stability requirement under aggregation consistent with the same underlying dynamics. M2 is tested in the D_3 coarse-graining sub-experiment in Section 7.2 (cross-level stability of weight rankings across token, n-gram, and sentence-cluster levels) and is distinct from the cross-domain test M5.

M3: Stability under recursive reuse. The third admissibility commitment. Weights must be tied to long-run contribution rather than momentary salience; a weight function that swings wildly under

small perturbations fails to track recursion and instead tracks noise. M3 is tested through cross-replicate stability metrics (in D_1, across the M = 20 replicate streams) and through bootstrap-CI procedures that quantify weight stability under within-domain resampling. The R2 cross-replicate Spearman threshold (0.7 minimum at calibration per Appendix B.4) is the operational form of M3 stability.

M4: Empirical derivability without exogenous semantics. The fourth admissibility commitment. There must exist at least one operational route by which w can be computed from observed data through a coherence proxy, without relying on interpretive judgments of meaning. M4 is satisfied by the derivation pipeline in *Formal Foundation of Induced Coherence Weights* (data → $\hat{C}$ → ρ → w) and by this paper's adaptation of that pipeline to the cross-domain validation architecture. M4 is the commitment that prevents coherence weighting from collapsing into hand-assigned semantic preference, and it is what makes the framework falsifiable rather than infinitely flexible.

M5: Domain-translation invariance. The fifth admissibility commitment, and the structurally distinct gate this paper carries through the validation architecture. If a domain mapping φ preserves coherence up to monotone transformation, then induced weights must be compatible under that mapping as well, again up to monotone transformation. M5 is not a fourth measurement signature alongside R1, R2, R3; it is an admissibility condition on the induced-w functional itself, evaluated across domains under matched estimator class. M5-pass licenses the eight-cell outcome matrix at full strength; M5-fail forecloses universality at the functional level and conditions the matrix's reading per the M5-fail interpretation matrix in Section 8.4. The operationalization in this paper computes the cross-domain ratio of mean-w over coherence-bearing features to mean-w over noise features, after rank-normalization, with the pre-registered threshold of factor-of-2 consistency across all domain pairs for at least 3 of 5 estimator classes.

M6: Recursive reinforcement compatibility. The sixth admissibility commitment. Induced weights must be compatible with the framework's commitment that coherence emerges through recursive reinforcement across scales; what persists is what reappears as enabling structure across iterations. M6 is satisfied by the induction pipeline's grounding in compression-based recoverability (Step C in the derivation pipeline), which defines relevance through recursive reinforcement detectable via compression. M6 does not enter the program's primary measurement layer as a

separately-evaluated condition; it is a structural commitment that the pipeline satisfies by construction.

Recursive coherence arbitration. The framework's mechanism by which reality is constituted. Structures persist when their coherence is repeatedly tested under perturbation and survives those tests; structures that cannot maintain coherence under arbitration are pruned. Reality is the cumulative result of arbitration across many iterations, not a static given. The arbitration mechanism is what authorizes Adaptive Realism's definition of the real and what makes persistence (rather than correspondence) the criterion of validity. R1 in this paper's program operationalizes the arbitration commitment at the measurement-operation level; induced w must predict which features survive perturbation (the arbitration's verdict) for the framework's apparatus to register as coherent with the framework's ontology.

Compression as validation. The third principle of Synthetic Epistemology. Epistemic value increases with representational economy; shorter, denser, more generative formulations are preferable when structurally equivalent. Compression is the operational signature of coherence; a system that compresses structurally equivalent formulations to denser representations is one whose internal logic tracks recursive reinforcement rather than surface variation. The induction pipeline in this paper grounds the validation operation in compression-based recoverability, making compression the empirical measurement layer through which the framework's coherence claims are tested. The K_1–K_5 estimator class spans different inductive biases about what compression means, and R2 tests whether the resulting induced-w distributions converge across those philosophically-decoupled compressors.

Pruning. The framework's mechanism for removing structure that has become incoherent. Pruning is licensed only when $\Delta C_S < 0 \wedge \frac{d\Omega_S}{dt} > 0$; the structure is reducing coherence and its removal expands the system's adaptive option space. Pruning is not failure; it is viability sorting. The framework's commitment to pruning extends to the framework's own components. Any module that resists pruning is not part of the system, and any extension to the corpus must be capable of self-collapse when it ceases to contribute to ΔC. This commitment is what makes the program in this paper structurally honest; each pre-registered failure mode is a configuration in which specific framework commitments would be pruned by the program's own outputs.

Volatility bound. The framework's safeguard against runaway destabilization. Coherence volatility σ_C must remain within a critical threshold $\theta_{\max}$; when σ_C exceeds $\theta_{\max}$, the system enters Tardigrade mode, a minimal-reactivity stabilization regime that freezes degradation and restores structural stability. The volatility bound is a regulatory constraint on adaptive responsiveness; too little adaptation registers as entropy accumulation, too much registers as brittleness. The bound does not enter the program's empirical measurements directly; it is a parameter of the framework's substrate-level dynamics and is referenced in the broader corpus's claims about adaptive limits.

Curvature constraint. The framework's constraint on the rate of structural change. Most systems cannot sustain structural change faster than approximately 2% per time cycle without fragmenting their feedback dynamics; the rate $r = (\frac{1}{C})(\frac{dC}{dt})$ must remain in approximately $[0.01, 0.03]$. The constraint is structural, not strategic; too fast registers as volatility, too slow registers as entropy accumulation. The curvature constraint is referenced in the broader corpus's claims about adaptive change rates and does not enter this paper's empirical measurements directly.

Drift = ΔC Loss. The fifth principle of Synthetic Epistemology. Failure in symbolic systems is defined as recursive coherence loss, contradiction under composition, or non-recoverable symbolic decay, not as deviation from external truth. The principle is the operational form of the framework's foreclosure of correspondence-truth. Validation does not test against an external standard, and failure does not mean failing to match an external standard. Failure means losing internal coherence under recursive operations, which the program's pre-registered outcome matrix names explicitly across the eight cells of within-domain failure and the three cells of M5-fail interpretation.

Meta-coherence. This paper's name for the framework's self-validation architecture. Meta-coherence is the framework's measurement operation cohering with the framework's own commitments across conditions the framework specifies in advance, with each result configuration bound to a specific consequence before any data is touched. The four conditions specified in Section 2, three within-domain measurement signatures (R1, R2, R3) plus one across-domain admissibility gate (M5), together operationalize meta-coherence as a falsifiable structure. Meta-coherence is not a substitute for empirical validation in the correspondence-truth sense; it is what validation looks like inside a framework that forecloses correspondence-truth at the ontological

level. The paper's contribution is the specification of meta-coherence at the level of an executable program.

Recursive coherence corpus. The body of theoretical work that this paper extends. The corpus operates by recursive layering; each paper compresses prior work to the level just above where the next paper begins. The relevant lineage for this paper consists of *Formal Foundations of Coherence Information Theory* (capacity and selective compression theorems conditional on bounded w), *Engineering Induced Coherence Weights for Coherence Information Theory* (engineering interface, benchmark suite, robustness harness), *Formal Foundation of Induced Coherence Weights* (derivation pipeline, M1–M6 admissibility commitments, residual hard seam at the choice of estimator), and the broader philosophical layer of Adaptive Realism, Synthetic Epistemology, and the diagnostic of symbolic lock-in. The next paper in the corpus is the empirical execution of the program specified here, which will take results back into theory and tighten or revise universality claims at whichever layer the data exposes.

Appendix E — Feature-Class Partition Lexicons

The M5 admissibility gate (Sections 2.5, 8.2, Appendix A.6, Appendix C.5) computes the cross-domain ratio of mean induced w over coherence-bearing features to mean induced w over noise features. The partition between coherence-bearing and noise classes within each domain is not derived from induced w; it is committed at pre-registration through documented sources whose epistemic status differs across domains in line with the transparency hierarchy. D_1's partition is exact by construction; D_2's partition is documented through biological literature; D_3's partition is curated through institutional-economics literature on Fed communication. This appendix specifies all three lexicons in full, with sources cited at the level of standing literature rather than at the level of specific publication endorsement.

E.1 D_1 partition: by construction

D_1's partition is exact. The generator's specification names which features carry coherence-bearing structure and which features are noise; the partition is a property of the generator, not of any analysis run on its outputs.

Coherence-bearing class:

- f_A — regime-conditional indicator (Property A)
- f_B — long-range coupling carrier with lag τ = 50 (Property B)
- f_{C_1} — first member of coalitional pair encoding (Z_t, s_t) (Property C)
- f_{C_2} — second member of coalitional pair encoding (Z_t, s_t) (Property C)
- f_D — drift carrier predicting regime transition (Property D)

Noise class:

- n_1 — uniform-high-entropy distractor
- n_2 — Zipf-structured distractor (α_{Zipf} locked at calibration)
- n_3 — low-frequency-dominant distractor (γ_{lowfreq} locked at calibration)

The partition is committed at pre-registration alongside the locked generator parameters from Appendix B. No post-hoc adjustment is permitted. The partition is uniform across all 20 replicate streams, all (K, A) cells in the 5×3 grid, and all stages of the analysis. Where R3 perturbations modify f_D (drift rate) or the regime transition matrix as structural interventions, the partition is unchanged; f_D remains in the coherence-bearing class regardless of perturbation magnitude, because the partition is specified by the feature's structural role in the generator, not by its post-perturbation behavior.

The exactness of D_1's partition is the asset Section 5.6 names. M5 evaluation in D_1 requires no inference about which features are coherence-bearing; the answer is in the generator's source code. The partition functions as the cleanest reference point in the cross-domain consistency check.

E.2 D_2 partition: by documented biology, per family

D_2's partition is committed at pre-registration per family, drawn from documented biology with explicit source attribution. The partition is bound by what the existing biological literature has established about each family's structural and functional positions, not by any inference the framework's apparatus would perform on the ortholog data.

Coherence-bearing class, per family. The class is the union of three position categories:

Category 1: Documented catalytic residues. For enzymatic families, positions annotated as catalytic in UniProt's "ACT_SITE" feature class and corroborated in the Catalytic Site Atlas (CSA, Thornton lab) where coverage exists. For non-enzymatic families with characterized active mechanisms, positions annotated in UniProt's "ACTIVE_SITE" or "BINDING" feature classes. Pre-registration commits to the UniProt release version and the CSA release version at the time of pre-registration; no version updates are absorbed during analysis.

Category 2: Fold-defining positions. Positions identified as load-bearing for the fold's structural integrity in published structural-biology literature. Sources for each family include: (a) primary structural studies of representative family members, identified through PDB depositions cross-referenced with the family's Pfam entry; (b) review articles on the family's structural class where they exist; (c) computational analyses identifying core residues through metrics such as residue depth, contact-order contribution, and packing density, drawn from tools such as PROSITE, SCOP/SCOPe domain definitions, and family-specific structural studies. Pre-registration commits to a per-family list of fold-defining positions with citations to the specific sources used, published as a supplementary table.

Category 3: Top conservation quintile. The 20% of positions with the highest conservation scores under the pre-registered conservation scoring method. Conservation scoring is committed at pre-registration as one of: ConSurf (Ben-Tal lab) at default parameters with the Bayesian scoring option, the Jensen-Shannon divergence method (Capra and Singh, 2007), or the Rate4Site method (Pupko et al., 2002). The specific method is locked at pre-registration; the choice of method is published with the pre-registration package.

Noise class, per family. The class is the union of two position categories:

Category 1: Surface-exposed non-functional residues. Positions with relative solvent accessibility above a pre-registered threshold (default 30% under the DSSP definition) AND not annotated to any functional role in UniProt's feature table. Functional role here covers ACT_SITE, BINDING, METAL, MOD_RES, DISULFID, INTRAMEM, MOTIF, REGION (where REGION specifies a functional sub-region rather than an arbitrary range), and SITE feature types.

Category 2: Bottom conservation quintile. The 20% of positions with the lowest conservation scores under the same conservation scoring method as Category 3 above.

Partition rules and edge cases.

A position appearing in both a coherence-bearing category and a noise category is excluded from the M5 computation for that family. Such positions are rare but possible (a position that is both surface-exposed and in the top conservation quintile, or both bottom-quintile-conserved and annotated as binding). Excluding ambiguous positions preserves the partition's epistemic cleanness without forcing arbitrary tiebreaking. The proportion of excluded positions per family is reported in the analysis and should fall below 5% by construction; families with higher exclusion rates are flagged for inspection without affecting the M5 verdict.

A position categorized into the coherence-bearing class through any of the three categories enters the class once; multiply-categorized positions (a documented catalytic residue that is also in the top conservation quintile, which is the typical case for true catalytic residues) are counted once.

The partition is committed per family with a specific Pfam family identifier, a specific UniProt release version, a specific CSA release version, a specific structural-source reference list, and a specific conservation scoring method. The committal is reproducible: independent re-derivation of the partition from the registered sources should produce identical results.

Sources at the level of standing literature.

The lexicon construction draws on the following bodies of work as reference frames:

- *Pfam database*: Mistry et al. on Pfam architecture and curation; Sonnhammer et al. on the original Pfam framework. The active Pfam release supplies the family selection and the multiple sequence alignments.
- *UniProt feature table*: The UniProt Consortium's documentation on feature types and annotation provenance. ACT_SITE, BINDING, METAL, MOD_RES, DISULFID, MOTIF, REGION, and SITE feature types are the operative annotations.
- *Catalytic Site Atlas*: Furnham et al. on the CSA's curation and scope. CSA entries provide curated catalytic-residue annotations corroborating UniProt where coverage exists.

- *Structural classifications*: SCOP (Murzin et al.), SCOPe (Fox et al.), and CATH (Orengo et al.) supply structural-class assignments and fold-level membership.
- *Conservation scoring*: ConSurf (Glaser et al., Ashkenazy et al.), Capra and Singh (2007) on Jensen-Shannon divergence for sequence conservation, Rate4Site (Pupko et al., 2002). Each represents a distinct philosophy of conservation scoring; the choice committed at pre-registration is published with the lexicon.
- *Solvent accessibility*: DSSP (Kabsch and Sander, 1983) on secondary structure assignment and accessible surface area; downstream tools NACCESS and FreeSASA implement the relative accessibility computation.

The references are listed at the level of source attribution; the pre-registration package cites specific papers and database release versions.

The discipline of partition commitment.

The partition is committed before analysis runs. Subsequent discoveries, that a particular position should have been included or excluded based on findings post-dating the pre-registration, that a different conservation scoring method would have produced cleaner results, that an alternative structural source contradicts the registered source, do not authorize partition modification. They authorize a separately registered re-analysis with a new partition committed in advance. The pre-registration discipline applies to the partition with the same force as it applies to the statistical thresholds and the calibration procedure.

E.3 D_3 FOMC token-class lexicon: institutional-economics curation

D_3's partition is committed through a curated lexicon of FOMC token classes, drawn from existing institutional-economics literature on Federal Reserve communication. The lexicon is published with the pre-registration package as a supplementary file and is not modified post-hoc.

The lexicon operates at the BPE Level 1 tokenization committed in Section 7.2. Coherence-bearing and noise classes are defined at the token level; multi-token phrases are decomposed into their constituent BPE tokens for the partition assignment.

Coherence-bearing class. The class is the union of four token categories drawn from documented FOMC communication patterns:

Category 1: Rate-decision tokens. Tokens that appear in the lexical formulation of federal funds target language. Examples include tokens decomposing the phrases "federal funds rate," "target range," "raise the target," "lower the target," "maintain the target," "basis points" (where the points

refer to the rate decision), and the numerical-token-and-modifier patterns associated with rate-level statements. The category is bounded by the operative rate-decision conventions across the 1994–2019 period; rate-decision language has been stable enough that the lexicon does not require period-specific subdivision within the primary corpus.

Category 2: Balance-sheet tokens. Tokens that appear in the lexical formulation of asset-purchase and balance-sheet-policy language. Examples include tokens decomposing the phrases "Treasury securities," "agency mortgage-backed securities," "agency debt," "asset purchases," "reinvestment," "balance sheet," "principal payments," "longer-term Treasury securities," and the operational-language tokens associated with QE1, QE2, QE3, the Maturity Extension Program, the gradual reduction of the balance sheet (2017–2019), and the System Open Market Account framing. The category is restricted to tokens that appear in policy-specifying contexts, not in tokens used for descriptive characterization of market conditions.

Category 3: Regime-shift indicators. Tokens that appear in the lexical formulation of policy framework changes. Examples include tokens decomposing the phrases "longer-run goals," "monetary policy strategy," "Statement on Longer-Run Goals and Monetary Policy Strategy," and the framework-specific tokens associated with the 2012 inflation-targeting framework (the "longer-run goal" formulation introduced at that point), the 2020 average-inflation-targeting framework (excluded from the primary corpus per the pandemic split, but included in the replication-set lexicon for completeness), and the framework-language tokens associated with documented framework transitions.

Category 4: Dual-mandate tokens. Tokens that appear in the lexical formulation of the Fed's statutory employment-and-price-stability mandate. Examples include tokens decomposing the phrases "maximum employment," "price stability," "moderate long-term interest rates," "dual mandate," "employment goals," "inflation objective," "longer-run goal of inflation," and tokens that appear in mandate-specifying contexts in FOMC statements. The category is restricted to tokens used in invocation of the mandate as an operative policy criterion, not tokens used in descriptive economic-condition characterization that happens to mention employment or inflation.

Noise class. The class is the union of four token categories that documented FOMC communication treats as procedural or conventional rather than substrate-policy-bearing:

Category 1: Conventional discourse markers. Tokens that appear in nearly every statement regardless of policy content. Examples include tokens decomposing connectives ("however," "moreover," "in addition," "furthermore"), framing markers ("Information received since the

Committee met"), and procedural connectives ("the Committee" used as an institutional-actor reference rather than as a policy-decision attribution).

Category 2: Procedural language. Tokens that describe how the meeting was conducted rather than what was decided. Examples include tokens decomposing the phrases "voting for," "voting against," "the Committee voted," "in support of," "the meeting concluded," and the procedural tokens associated with the Federal Reserve System's institutional protocol descriptions.

Category 3: Calendar and date tokens. Tokens that decompose date references, meeting-cycle references, and time-stamp language. Examples include month names, year tokens, day tokens, and the tokens decomposing phrases such as "the next meeting," "this meeting," "since the previous meeting", language that supplies temporal indexing without policy-substrate content.

Category 4: Committee-membership references. Tokens that identify meeting attendees, voting alignments at the procedural level, and Federal Reserve Bank presidents' names where they appear in voting-attribution contexts. The category is bounded to procedural attribution; tokens identifying chair statements that subsequently quote substrate-policy language are not in the noise class because the quoted content is itself policy-substantive.

Partition rules and edge cases.

A token appearing in both a coherence-bearing category and a noise category is excluded from the M5 computation for D_3. Such tokens are rare but possible (a token that decomposes a date reference but also appears as the year-token in a regime-shift indicator). Excluding ambiguous tokens preserves the partition's epistemic cleanness.

The lexicon is closed at pre-registration. Subsequent discoveries, that a particular token should have been included in either class based on findings post-dating the pre-registration, that the lexicon undercovers or overcovers a specific category, do not authorize lexicon modification. They authorize a separately registered re-analysis with a new lexicon committed in advance.

Sources at the level of standing literature.

The lexicon construction draws on the following bodies of work as reference frames for what FOMC communication treats as substrate-policy-bearing versus procedural or conventional:

- *Fed-communication transparency literature*: Blinder, Ehrmann, Fratzscher, de Haan, and Jansen on central-bank communication transparency and effectiveness; Ehrmann and Fratzscher on FOMC verbal-communication patterns; Hansen, McMahon, and Prat on transparency and FOMC deliberation.

- *Forward-guidance language analysis*: Campbell, Evans, Fisher, and Justiniano on forward guidance and its lexical conventions; Del Negro, Giannoni, and Patterson on forward-guidance puzzles and the language conventions used to convey them; Bundick and Smith on Fed-communication and yield-curve responses.
- *Textual analysis of FOMC communication*: Hansen and McMahon on FOMC text mining; Husted, Rogers, and Sun on Fed-communication text analysis; Acosta on FOMC sentiment and content analysis; Cieslak and Vissing-Jorgensen on Fed-communication and risk premia.
- *Policy-framework transitions*: Goodfriend on the inflation-targeting framework's introduction; Bernanke, Reinhart, and Sack on the language conventions accompanying nontraditional monetary policy; Yellen's 2017 speech on policy-framework evolution; the 2012 *Statement on Longer-Run Goals and Monetary Policy Strategy* itself.
- *Balance-sheet policy language*: Krishnamurthy and Vissing-Jorgensen on the language of large-scale asset purchases; D'Amico, English, López-Salido, and Nelson on Federal Reserve balance-sheet operations and their lexical conventions.
- *Linguistic-convention shifts*: documented historiographic analyses of the "considerable period" introduction (August 2003), the "patient" introduction (2014), the "extended period" formulation, and the forward-guidance threshold language (2012). Sources include Federal Reserve Board archival materials cross-referenced with monetary-policy-history scholarship.

The references are listed at the level of source attribution; the pre-registration package cites specific papers, specific FOMC archival materials, and the specific lexicon entries each source supports.

The lexicon's epistemic status.

The lexicon is not a ground-truth annotation of FOMC statements. It is a documented compression of FOMC communication, drawn from the institutional-economics literature's accumulated work on what Fed communication treats as policy-substantive versus procedural. The framework treats the lexicon exactly as it treats Pfam in D_2, as a documented compression from a community of researchers whose inductive biases differ from the framework's own, used as an external partition reference rather than as a verdict on what induced *w* should produce. Convergence between induced w and the lexicon's coherence-bearing class is one meta-coherence signal; divergence is informative. The lexicon supplies the partition; the framework's apparatus produces induced w; the M5 ratio compares the two without privileging either.

The lexicon's published form includes for each entry: the token (in BPE form), the assigned class (coherence-bearing or noise), the category within the class (rate-decision, balance-sheet, regime-shift, dual-mandate, conventional-discourse, procedural, calendar-date, or committee-membership), the source citation supporting the assignment, and the rationale where the source citation alone is insufficient. The published form is approximately 800–1,200 tokens in the coherence-bearing class and 1,500–2,500 tokens in the noise class after BPE decomposition; the remaining vocabulary tokens are unclassified and excluded from the M5 ratio computation.

E.4 The cross-domain status of the partition lexicons

The three partition lexicons differ in epistemic status by design. D_1's is exact; D_2's is documented-biological; D_3's is curated-institutional. The transparency hierarchy in Section 4 maps directly onto the partition's epistemic status at each tier, and the cross-domain consistency check at the M5 gate operates across the three tiers without requiring uniformity in how the partitions are derived.

The framework's commitment is not that the three partitions are equally certain. The framework's commitment is that each partition is bound by the best available compression at its domain's transparency tier, that none of the partitions depends on induced w, and that the M5 admissibility ratio is a meaningful cross-domain comparison precisely because the partitions are derived from sources external to the framework's apparatus. The structural transparency gradient runs through the partition lexicons themselves, and the M5 verdict's interpretation accommodates that gradient; a strong M5-pass across all three tiers registers as substrate-transparency-invariance of the induced-*w* functional, which is the strongest universality claim available under the framework's own ontology.

The lexicons in this appendix are the operational form of the framework's commitment to compression-with-compression comparison rather than compression-against-truth comparison. Each lexicon is a compression of the substrate's structural distinctions through the inductive biases of a particular community (the generator's writer in D_1, biology's curatorial community in D_2, institutional-economics scholarship in D_3), and the M5 ratio asks whether the framework's induced-*w* compression produces consistent structural distinctions when read against those reference compressions.